

| | | | | | | | | | | | | | | |
|----|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| 40 | 0,0013 | 0,0063 | 0,0126 | 0,0315 | 0,1903 | 0,3208 | 0,6807 | 1,1673 | 1,6839 | 2,0211 | 2,3289 | 2,7045 | 2,9712 | 3,5510 |
| 41 | 0,0013 | 0,0063 | 0,0126 | 0,0315 | 0,1903 | 0,3208 | 0,6805 | 1,1669 | 1,6829 | 2,0195 | 2,3267 | 2,7012 | 2,9670 | 3,5443 |
| 42 | 0,0013 | 0,0063 | 0,0126 | 0,0315 | 0,1903 | 0,3207 | 0,6804 | 1,1665 | 1,6820 | 2,0181 | 2,3246 | 2,6981 | 2,9630 | 3,5377 |
| 43 | 0,0013 | 0,0063 | 0,0126 | 0,0315 | 0,1903 | 0,3207 | 0,6802 | 1,1661 | 1,6811 | 2,0167 | 2,3226 | 2,6951 | 2,9592 | 3,5316 |
| 44 | 0,0013 | 0,0063 | 0,0126 | 0,0315 | 0,1902 | 0,3206 | 0,6801 | 1,1657 | 1,6802 | 2,0154 | 2,3207 | 2,6923 | 2,9555 | 3,5258 |
| 45 | 0,0013 | 0,0063 | 0,0126 | 0,0315 | 0,1902 | 0,3206 | 0,6800 | 1,1654 | 1,6794 | 2,0141 | 2,3189 | 2,6896 | 2,9521 | 3,5203 |
| 46 | 0,0013 | 0,0063 | 0,0126 | 0,0315 | 0,1902 | 0,3206 | 0,6799 | 1,1651 | 1,6787 | 2,0129 | 2,3172 | 2,6870 | 2,9488 | 3,5149 |
| 47 | 0,0013 | 0,0063 | 0,0126 | 0,0315 | 0,1902 | 0,3205 | 0,6797 | 1,1647 | 1,6779 | 2,0117 | 2,3155 | 2,6846 | 2,9456 | 3,5099 |
| 48 | 0,0013 | 0,0063 | 0,0126 | 0,0315 | 0,1901 | 0,3205 | 0,6796 | 1,1644 | 1,6772 | 2,0106 | 2,3139 | 2,6822 | 2,9426 | 3,5050 |
| 49 | 0,0013 | 0,0063 | 0,0126 | 0,0315 | 0,1901 | 0,3204 | 0,6795 | 1,1642 | 1,6766 | 2,0096 | 2,3124 | 2,6800 | 2,9397 | 3,5005 |
| 50 | 0,0013 | 0,0063 | 0,0126 | 0,0315 | 0,1901 | 0,3204 | 0,6794 | 1,1639 | 1,6759 | 2,0086 | 2,3109 | 2,6778 | 2,9370 | 3,4960 |

Bibliografia.

Castañeda, DFN. “Estatística Econômica” Site: <http://www.geocities.com/danielneyra/aee.zip>

Castañeda, DFN. “Demografia” Site: <http://www.geocities.com/danielneyra/demog.zip>

Castañeda, DFN. “Séries Temporais” Site: <http://www.geocities.com/danielneyra/Seriest.zip>

Anexo 3- Valores para a distribuição t

| Nível de significância (bilateral) | | | | | | | | | | | | | | |
|------------------------------------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| gl | 0,999 | 0,995 | 0,990 | 0,975 | 0,850 | 0,750 | 0,500 | 0,250 | 0,100 | 0,050 | 0,025 | 0,010 | 0,005 | 0,001 |
| 1 | 0,0016 | 0,0079 | 0,0157 | 0,0393 | 0,2401 | 0,4142 | 1,0000 | 2,4142 | 6,3137 | 12,706 | 25,452 | 63,656 | 127,32 | 636,58 |
| 2 | 0,0014 | 0,0071 | 0,0141 | 0,0354 | 0,2146 | 0,3651 | 0,8165 | 1,6036 | 2,9200 | 4,3027 | 6,2054 | 9,9250 | 14,09 | 31,60 |
| 3 | 0,0014 | 0,0068 | 0,0136 | 0,0340 | 0,2060 | 0,3492 | 0,7649 | 1,4226 | 2,3534 | 3,1824 | 4,1765 | 5,8408 | 7,4532 | 12,924 |
| 4 | 0,0013 | 0,0067 | 0,0133 | 0,0333 | 0,2017 | 0,3414 | 0,7407 | 1,3444 | 2,1318 | 2,7765 | 3,4954 | 4,6041 | 5,5975 | 8,6101 |
| 5 | 0,0013 | 0,0066 | 0,0132 | 0,0329 | 0,1991 | 0,3367 | 0,7267 | 1,3009 | 2,0150 | 2,5706 | 3,1634 | 4,0321 | 4,7733 | 6,8685 |
| 6 | 0,0013 | 0,0065 | 0,0131 | 0,0327 | 0,1974 | 0,3336 | 0,7176 | 1,2733 | 1,9432 | 2,4469 | 2,9687 | 3,7074 | 4,3168 | 5,958 |
| 7 | 0,0013 | 0,0065 | 0,0130 | 0,0325 | 0,1962 | 0,3315 | 0,7111 | 1,2543 | 1,8946 | 2,3646 | 2,8412 | 3,4995 | 4,0294 | 5,4081 |
| 8 | 0,0013 | 0,0065 | 0,0129 | 0,0323 | 0,1953 | 0,3298 | 0,7064 | 1,2403 | 1,8595 | 2,3060 | 2,7515 | 3,3554 | 3,8325 | 5,0414 |
| 9 | 0,0013 | 0,0064 | 0,0129 | 0,0322 | 0,1946 | 0,3286 | 0,7027 | 1,2297 | 1,8331 | 2,2622 | 2,6850 | 3,2498 | 3,6896 | 4,7809 |
| 10 | 0,0013 | 0,0064 | 0,0129 | 0,0321 | 0,1941 | 0,3276 | 0,6998 | 1,2213 | 1,8125 | 2,2281 | 2,6338 | 3,1693 | 3,5814 | 4,5868 |
| 11 | 0,0013 | 0,0064 | 0,0128 | 0,0321 | 0,1936 | 0,3267 | 0,6974 | 1,2145 | 1,7959 | 2,2010 | 2,5931 | 3,1058 | 3,4966 | 4,4369 |
| 12 | 0,0013 | 0,0064 | 0,0128 | 0,0320 | 0,1932 | 0,3261 | 0,6955 | 1,2089 | 1,7823 | 2,1788 | 2,5600 | 3,0545 | 3,4284 | 4,3178 |
| 13 | 0,0013 | 0,0064 | 0,0128 | 0,0319 | 0,1929 | 0,3255 | 0,6938 | 1,2041 | 1,7709 | 2,1604 | 2,5326 | 3,0123 | 3,3725 | 4,2209 |
| 14 | 0,0013 | 0,0064 | 0,0128 | 0,0319 | 0,1926 | 0,3250 | 0,6924 | 1,2001 | 1,7613 | 2,1448 | 2,5096 | 2,9768 | 3,3257 | 4,1403 |
| 15 | 0,0013 | 0,0064 | 0,0127 | 0,0319 | 0,1924 | 0,3246 | 0,6912 | 1,1967 | 1,7531 | 2,1315 | 2,4899 | 2,9467 | 3,2860 | 4,0728 |
| 16 | 0,0013 | 0,0064 | 0,0127 | 0,0318 | 0,1922 | 0,3242 | 0,6901 | 1,1937 | 1,7459 | 2,1199 | 2,4729 | 2,9208 | 3,2520 | 4,0149 |
| 17 | 0,0013 | 0,0064 | 0,0127 | 0,0318 | 0,1920 | 0,3239 | 0,6892 | 1,1910 | 1,7396 | 2,1098 | 2,4581 | 2,8982 | 3,2224 | 3,9651 |
| 18 | 0,0013 | 0,0064 | 0,0127 | 0,0318 | 0,1919 | 0,3236 | 0,6884 | 1,1887 | 1,7341 | 2,1009 | 2,4450 | 2,8784 | 3,1966 | 3,9217 |
| 19 | 0,0013 | 0,0063 | 0,0127 | 0,0318 | 0,1917 | 0,3233 | 0,6876 | 1,1866 | 1,7291 | 2,0930 | 2,4334 | 2,8609 | 3,1737 | 3,8833 |
| 20 | 0,0013 | 0,0063 | 0,0127 | 0,0317 | 0,1916 | 0,3231 | 0,6870 | 1,1848 | 1,7247 | 2,0860 | 2,4231 | 2,8453 | 3,1534 | 3,8496 |
| 21 | 0,0013 | 0,0063 | 0,0127 | 0,0317 | 0,1915 | 0,3229 | 0,6864 | 1,1831 | 1,7207 | 2,0796 | 2,4138 | 2,8314 | 3,1352 | 3,8193 |
| 22 | 0,0013 | 0,0063 | 0,0127 | 0,0317 | 0,1914 | 0,3227 | 0,6858 | 1,1815 | 1,7171 | 2,0739 | 2,4055 | 2,8188 | 3,1188 | 3,7922 |
| 23 | 0,0013 | 0,0063 | 0,0127 | 0,0317 | 0,1913 | 0,3225 | 0,6853 | 1,1802 | 1,7139 | 2,0687 | 2,3979 | 2,8073 | 3,1040 | 3,7676 |
| 24 | 0,0013 | 0,0063 | 0,0127 | 0,0317 | 0,1912 | 0,3223 | 0,6848 | 1,1789 | 1,7109 | 2,0639 | 2,3910 | 2,7970 | 3,0905 | 3,7454 |
| 25 | 0,0013 | 0,0063 | 0,0127 | 0,0317 | 0,1911 | 0,3222 | 0,6844 | 1,1777 | 1,7081 | 2,0595 | 2,3846 | 2,7874 | 3,0782 | 3,7251 |
| 26 | 0,0013 | 0,0063 | 0,0127 | 0,0316 | 0,1910 | 0,3220 | 0,6840 | 1,1766 | 1,7056 | 2,0555 | 2,3788 | 2,7787 | 3,0669 | 3,7067 |
| 27 | 0,0013 | 0,0063 | 0,0126 | 0,0316 | 0,1909 | 0,3219 | 0,6837 | 1,1756 | 1,7033 | 2,0518 | 2,3734 | 2,7707 | 3,0565 | 3,6895 |
| 28 | 0,0013 | 0,0063 | 0,0126 | 0,0316 | 0,1909 | 0,3218 | 0,6834 | 1,1747 | 1,7011 | 2,0484 | 2,3685 | 2,7633 | 3,0470 | 3,6739 |
| 29 | 0,0013 | 0,0063 | 0,0126 | 0,0316 | 0,1908 | 0,3217 | 0,6830 | 1,1739 | 1,6991 | 2,0452 | 2,3638 | 2,7564 | 3,0380 | 3,6595 |
| 30 | 0,0013 | 0,0063 | 0,0126 | 0,0316 | 0,1908 | 0,3216 | 0,6828 | 1,1731 | 1,6973 | 2,0423 | 2,3596 | 2,7500 | 3,0298 | 3,6460 |
| 31 | 0,0013 | 0,0063 | 0,0126 | 0,0316 | 0,1907 | 0,3215 | 0,6825 | 1,1723 | 1,6955 | 2,0395 | 2,3556 | 2,7440 | 3,0221 | 3,6335 |
| 32 | 0,0013 | 0,0063 | 0,0126 | 0,0316 | 0,1907 | 0,3214 | 0,6822 | 1,1716 | 1,6939 | 2,0369 | 2,3518 | 2,7385 | 3,0149 | 3,6218 |
| 33 | 0,0013 | 0,0063 | 0,0126 | 0,0316 | 0,1906 | 0,3213 | 0,6820 | 1,1710 | 1,6924 | 2,0345 | 2,3483 | 2,7333 | 3,0082 | 3,6109 |
| 34 | 0,0013 | 0,0063 | 0,0126 | 0,0316 | 0,1906 | 0,3212 | 0,6818 | 1,1703 | 1,6909 | 2,0322 | 2,3451 | 2,7284 | 3,0020 | 3,6007 |
| 35 | 0,0013 | 0,0063 | 0,0126 | 0,0316 | 0,1905 | 0,3212 | 0,6816 | 1,1698 | 1,6896 | 2,0301 | 2,3420 | 2,7238 | 2,9961 | 3,5911 |
| 36 | 0,0013 | 0,0063 | 0,0126 | 0,0316 | 0,1905 | 0,3211 | 0,6814 | 1,1692 | 1,6883 | 2,0281 | 2,3391 | 2,7195 | 2,9905 | 3,5821 |
| 37 | 0,0013 | 0,0063 | 0,0126 | 0,0316 | 0,1904 | 0,3210 | 0,6812 | 1,1687 | 1,6871 | 2,0262 | 2,3363 | 2,7154 | 2,9853 | 3,5737 |
| 38 | 0,0013 | 0,0063 | 0,0126 | 0,0315 | 0,1904 | 0,3210 | 0,6810 | 1,1682 | 1,6860 | 2,0244 | 2,3337 | 2,7116 | 2,9803 | 3,5657 |
| 39 | 0,0013 | 0,0063 | 0,0126 | 0,0315 | 0,1904 | 0,3209 | 0,6808 | 1,1677 | 1,6849 | 2,0227 | 2,3313 | 2,7079 | 2,9756 | 3,5581 |

| | | | | | | | | | | | | | | |
|----|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| 42 | 19,238 | 22,138 | 23,650 | 25,999 | 28,144 | 30,765 | 35,510 | 47,766 | 54,090 | 58,124 | 61,777 | 66,206 | 69,336 | 76,084 |
| 43 | 19,905 | 22,860 | 24,398 | 26,785 | 28,965 | 31,625 | 36,436 | 48,840 | 55,230 | 59,304 | 62,990 | 67,459 | 70,616 | 77,418 |
| 44 | 20,576 | 23,584 | 25,148 | 27,575 | 29,787 | 32,487 | 37,363 | 49,913 | 56,369 | 60,481 | 64,201 | 68,710 | 71,892 | 78,749 |
| 45 | 21,251 | 24,311 | 25,901 | 28,366 | 30,612 | 33,350 | 38,291 | 50,985 | 57,505 | 61,656 | 65,410 | 69,957 | 73,166 | 80,078 |
| 46 | 21,929 | 25,041 | 26,657 | 29,160 | 31,439 | 34,215 | 39,220 | 52,056 | 58,641 | 62,830 | 66,616 | 71,201 | 74,437 | 81,400 |
| 47 | 22,610 | 25,775 | 27,416 | 29,956 | 32,268 | 35,081 | 40,149 | 53,127 | 59,774 | 64,001 | 67,821 | 72,443 | 75,704 | 82,720 |
| 48 | 23,294 | 26,511 | 28,177 | 30,754 | 33,098 | 35,949 | 41,079 | 54,196 | 60,907 | 65,171 | 69,023 | 73,683 | 76,969 | 84,037 |
| 49 | 23,983 | 27,249 | 28,941 | 31,555 | 33,930 | 36,818 | 42,010 | 55,265 | 62,038 | 66,339 | 70,222 | 74,919 | 78,231 | 85,350 |
| 50 | 24,674 | 27,991 | 29,707 | 32,357 | 34,764 | 37,689 | 42,942 | 56,334 | 63,167 | 67,505 | 71,420 | 76,154 | 79,490 | 86,660 |

Anexo 2- Valores para a distribuição Qui Quadrado

| <i>gl</i> | 0,999 | 0,995 | 0,990 | 0,975 | 0,950 | 0,900 | 0,750 | 0,250 | 0,100 | 0,050 | 0,025 | 0,010 | 0,005 | 0,001 |
|-----------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| 1 | 0,000 | 0,000 | 0,000 | 0,001 | 0,004 | 0,016 | 0,102 | 1,323 | 2,706 | 3,841 | 5,024 | 6,635 | 7,879 | 10,827 |
| 2 | 0,002 | 0,010 | 0,020 | 0,051 | 0,103 | 0,211 | 0,575 | 2,773 | 4,605 | 5,991 | 7,378 | 9,210 | 10,597 | 13,815 |
| 3 | 0,024 | 0,072 | 0,115 | 0,216 | 0,352 | 0,584 | 1,213 | 4,108 | 6,251 | 7,815 | 9,348 | 11,345 | 12,838 | 16,266 |
| 4 | 0,091 | 0,207 | 0,297 | 0,484 | 0,711 | 1,064 | 1,923 | 5,385 | 7,779 | 9,488 | 11,143 | 13,277 | 14,860 | 18,466 |
| 5 | 0,210 | 0,412 | 0,554 | 0,831 | 1,145 | 1,610 | 2,675 | 6,626 | 9,236 | 11,070 | 12,832 | 15,086 | 16,750 | 20,515 |
| 6 | 0,381 | 0,676 | 0,872 | 1,237 | 1,635 | 2,204 | 3,455 | 7,841 | 10,645 | 12,592 | 14,449 | 16,812 | 18,548 | 22,457 |
| 7 | 0,599 | 0,989 | 1,239 | 1,690 | 2,167 | 2,833 | 4,255 | 9,037 | 12,017 | 14,067 | 16,013 | 18,475 | 20,278 | 24,321 |
| 8 | 0,857 | 1,344 | 1,647 | 2,180 | 2,733 | 3,490 | 5,071 | 10,219 | 13,362 | 15,507 | 17,535 | 20,090 | 21,955 | 26,124 |
| 9 | 1,152 | 1,735 | 2,088 | 2,700 | 3,325 | 4,168 | 5,899 | 11,389 | 14,684 | 16,919 | 19,023 | 21,666 | 23,589 | 27,877 |
| 10 | 1,479 | 2,156 | 2,558 | 3,247 | 3,940 | 4,865 | 6,737 | 12,549 | 15,987 | 18,307 | 20,483 | 23,209 | 25,188 | 29,588 |
| 11 | 1,834 | 2,603 | 3,053 | 3,816 | 4,575 | 5,578 | 7,584 | 13,701 | 17,275 | 19,675 | 21,920 | 24,725 | 26,757 | 31,264 |
| 12 | 2,214 | 3,074 | 3,571 | 4,404 | 5,226 | 6,304 | 8,438 | 14,845 | 18,549 | 21,026 | 23,337 | 26,217 | 28,300 | 32,909 |
| 13 | 2,617 | 3,565 | 4,107 | 5,009 | 5,892 | 7,041 | 9,299 | 15,984 | 19,812 | 22,362 | 24,736 | 27,688 | 29,819 | 34,527 |
| 14 | 3,041 | 4,075 | 4,660 | 5,629 | 6,571 | 7,790 | 10,165 | 17,117 | 21,064 | 23,685 | 26,119 | 29,141 | 31,319 | 36,124 |
| 15 | 3,483 | 4,601 | 5,229 | 6,262 | 7,261 | 8,547 | 11,037 | 18,245 | 22,307 | 24,996 | 27,488 | 30,578 | 32,801 | 37,698 |
| 16 | 3,942 | 5,142 | 5,812 | 6,908 | 7,962 | 9,312 | 11,912 | 19,369 | 23,542 | 26,296 | 28,845 | 32,000 | 34,267 | 39,252 |
| 17 | 4,416 | 5,697 | 6,408 | 7,564 | 8,672 | 10,085 | 12,792 | 20,489 | 24,769 | 27,587 | 30,191 | 33,409 | 35,718 | 40,791 |
| 18 | 4,905 | 6,265 | 7,015 | 8,231 | 9,390 | 10,865 | 13,675 | 21,605 | 25,989 | 28,869 | 31,526 | 34,805 | 37,156 | 42,312 |
| 19 | 5,407 | 6,844 | 7,633 | 8,907 | 10,117 | 11,651 | 14,562 | 22,718 | 27,204 | 30,144 | 32,852 | 36,191 | 38,582 | 43,819 |
| 20 | 5,921 | 7,434 | 8,260 | 9,591 | 10,851 | 12,443 | 15,452 | 23,828 | 28,412 | 31,410 | 34,170 | 37,566 | 39,997 | 45,314 |
| 21 | 6,447 | 8,034 | 8,897 | 10,283 | 11,591 | 13,240 | 16,344 | 24,935 | 29,615 | 32,671 | 35,479 | 38,932 | 41,401 | 46,796 |
| 22 | 6,983 | 8,643 | 9,542 | 10,982 | 12,338 | 14,041 | 17,240 | 26,039 | 30,813 | 33,924 | 36,781 | 40,289 | 42,796 | 48,268 |
| 23 | 7,529 | 9,260 | 10,196 | 11,689 | 13,091 | 14,848 | 18,137 | 27,141 | 32,007 | 35,172 | 38,076 | 41,638 | 44,181 | 49,728 |
| 24 | 8,085 | 9,886 | 10,856 | 12,401 | 13,848 | 15,659 | 19,037 | 28,241 | 33,196 | 36,415 | 39,364 | 42,980 | 45,558 | 51,179 |
| 25 | 8,649 | 10,520 | 11,524 | 13,120 | 14,611 | 16,473 | 19,939 | 29,339 | 34,382 | 37,652 | 40,646 | 44,314 | 46,928 | 52,619 |
| 26 | 9,222 | 11,160 | 12,198 | 13,844 | 15,379 | 17,292 | 20,843 | 30,435 | 35,563 | 38,885 | 41,923 | 45,642 | 48,290 | 54,051 |
| 27 | 9,803 | 11,808 | 12,878 | 14,573 | 16,151 | 18,114 | 21,749 | 31,528 | 36,741 | 40,113 | 43,195 | 46,963 | 49,645 | 55,475 |
| 28 | 10,391 | 12,461 | 13,565 | 15,308 | 16,928 | 18,939 | 22,657 | 32,620 | 37,916 | 41,337 | 44,461 | 48,278 | 50,994 | 56,892 |
| 29 | 10,986 | 13,121 | 14,256 | 16,047 | 17,708 | 19,768 | 23,567 | 33,711 | 39,087 | 42,557 | 45,722 | 49,588 | 52,335 | 58,301 |
| 30 | 11,588 | 13,787 | 14,953 | 16,791 | 18,493 | 20,599 | 24,478 | 34,800 | 40,256 | 43,773 | 46,979 | 50,892 | 53,672 | 59,702 |
| 31 | 12,196 | 14,458 | 15,655 | 17,539 | 19,281 | 21,434 | 25,390 | 35,887 | 41,422 | 44,985 | 48,232 | 52,191 | 55,002 | 61,098 |
| 32 | 12,810 | 15,134 | 16,362 | 18,291 | 20,072 | 22,271 | 26,304 | 36,973 | 42,585 | 46,194 | 49,480 | 53,486 | 56,328 | 62,487 |
| 33 | 13,431 | 15,815 | 17,073 | 19,047 | 20,867 | 23,110 | 27,219 | 38,058 | 43,745 | 47,400 | 50,725 | 54,775 | 57,648 | 63,869 |
| 34 | 14,057 | 16,501 | 17,789 | 19,806 | 21,664 | 23,952 | 28,136 | 39,141 | 44,903 | 48,602 | 51,966 | 56,061 | 58,964 | 65,247 |
| 35 | 14,688 | 17,192 | 18,509 | 20,569 | 22,465 | 24,797 | 29,054 | 40,223 | 46,059 | 49,802 | 53,203 | 57,342 | 60,275 | 66,619 |
| 36 | 15,324 | 17,887 | 19,233 | 21,336 | 23,269 | 25,643 | 29,973 | 41,304 | 47,212 | 50,998 | 54,437 | 58,619 | 61,581 | 67,985 |
| 37 | 15,965 | 18,586 | 19,960 | 22,106 | 24,075 | 26,492 | 30,893 | 42,383 | 48,363 | 52,192 | 55,668 | 59,893 | 62,883 | 69,348 |
| 38 | 16,611 | 19,289 | 20,691 | 22,878 | 24,884 | 27,343 | 31,815 | 43,462 | 49,513 | 53,384 | 56,895 | 61,162 | 64,181 | 70,704 |
| 39 | 17,261 | 19,996 | 21,426 | 23,654 | 25,695 | 28,196 | 32,737 | 44,539 | 50,660 | 54,572 | 58,120 | 62,428 | 65,475 | 72,055 |
| 40 | 17,917 | 20,707 | 22,164 | 24,433 | 26,509 | 29,051 | 33,660 | 45,616 | 51,805 | 55,758 | 59,342 | 63,691 | 66,766 | 73,403 |
| 41 | 18,576 | 21,421 | 22,906 | 25,215 | 27,326 | 29,907 | 34,585 | 46,692 | 52,949 | 56,942 | 60,561 | 64,950 | 68,053 | 74,744 |

Anexo 1- Tabela para a distribuição Normal - integral compreendida entre 0 e z

| | 0,00 | 0,01 | 0,02 | 0,03 | 0,04 | 0,05 | 0,06 | 0,07 | 0,08 | 0,09 |
|-----|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| 0,0 | 0,0000000 | 0,0039894 | 0,0079784 | 0,0119665 | 0,0159535 | 0,0199389 | 0,0239223 | 0,0279032 | 0,0318814 | 0,0358565 |
| 0,1 | 0,0398279 | 0,0437954 | 0,0477585 | 0,0517168 | 0,0556700 | 0,0596177 | 0,0635595 | 0,0674949 | 0,0714237 | 0,0753454 |
| 0,2 | 0,0792597 | 0,0831661 | 0,0870644 | 0,0909541 | 0,0948348 | 0,0987063 | 0,1025681 | 0,1064198 | 0,1102612 | 0,1140918 |
| 0,3 | 0,1179114 | 0,1217195 | 0,1255158 | 0,1293000 | 0,1330717 | 0,1368306 | 0,1405764 | 0,1443087 | 0,1480272 | 0,1517317 |
| 0,4 | 0,1554217 | 0,1590970 | 0,1627572 | 0,1664021 | 0,1700314 | 0,1736448 | 0,1772419 | 0,1808225 | 0,1843863 | 0,1879331 |
| 0,5 | 0,1914625 | 0,1949743 | 0,1984682 | 0,2019441 | 0,2054015 | 0,2088403 | 0,2122603 | 0,2156612 | 0,2190427 | 0,2224047 |
| 0,6 | 0,2257469 | 0,2290692 | 0,2323712 | 0,2356528 | 0,2389138 | 0,2421540 | 0,2453732 | 0,2485712 | 0,2517478 | 0,2549030 |
| 0,7 | 0,2580364 | 0,2611480 | 0,2642376 | 0,2673050 | 0,2703501 | 0,2733727 | 0,2763728 | 0,2793501 | 0,2823046 | 0,2852362 |
| 0,8 | 0,2881447 | 0,2910300 | 0,2938920 | 0,2967307 | 0,2995459 | 0,3023375 | 0,3051055 | 0,3078498 | 0,3105704 | 0,3132671 |
| 0,9 | 0,3159399 | 0,3185888 | 0,3212136 | 0,3238145 | 0,3263912 | 0,3289439 | 0,3314724 | 0,3339768 | 0,3364569 | 0,3389129 |
| 1,0 | 0,3413447 | 0,3437523 | 0,3461358 | 0,3484950 | 0,3508300 | 0,3531409 | 0,3554277 | 0,3576903 | 0,3599289 | 0,3621434 |
| 1,1 | 0,3643339 | 0,3665004 | 0,3686431 | 0,3707618 | 0,3728568 | 0,3749280 | 0,3769755 | 0,3789995 | 0,3809998 | 0,3829767 |
| 1,2 | 0,3849303 | 0,3868605 | 0,3887675 | 0,3906514 | 0,3925122 | 0,3943502 | 0,3961653 | 0,3979576 | 0,3997274 | 0,4014746 |
| 1,3 | 0,4031995 | 0,4049020 | 0,4065824 | 0,4082408 | 0,4098773 | 0,4114919 | 0,4130850 | 0,4146565 | 0,4162066 | 0,4177355 |
| 1,4 | 0,4192433 | 0,4207301 | 0,4221961 | 0,4236414 | 0,4250663 | 0,4264707 | 0,4278549 | 0,4292191 | 0,4305633 | 0,4318879 |
| 1,5 | 0,4331928 | 0,4344783 | 0,4357445 | 0,4369916 | 0,4382198 | 0,4394292 | 0,4406200 | 0,4417924 | 0,4429466 | 0,4440826 |
| 1,6 | 0,4452007 | 0,4463011 | 0,4473839 | 0,4484493 | 0,4494974 | 0,4505285 | 0,4515428 | 0,4525403 | 0,4535214 | 0,4544861 |
| 1,7 | 0,4554346 | 0,4563671 | 0,4572838 | 0,4581849 | 0,4590705 | 0,4599409 | 0,4607961 | 0,4616365 | 0,4624621 | 0,4632731 |
| 1,8 | 0,4640697 | 0,4648522 | 0,4656206 | 0,4663751 | 0,4671159 | 0,4678433 | 0,4685573 | 0,4692582 | 0,4699460 | 0,4706211 |
| 1,9 | 0,4712835 | 0,4719335 | 0,4725711 | 0,4731967 | 0,4738102 | 0,4744120 | 0,4750022 | 0,4755809 | 0,4761483 | 0,4767046 |
| 2,0 | 0,4772499 | 0,4777845 | 0,4783084 | 0,4788218 | 0,4793249 | 0,4798179 | 0,4803008 | 0,4807739 | 0,4812373 | 0,4816912 |
| 2,1 | 0,4821356 | 0,4825709 | 0,4829970 | 0,4834143 | 0,4838227 | 0,4842224 | 0,4846137 | 0,4849966 | 0,4853713 | 0,4857379 |
| 2,2 | 0,4860966 | 0,4864475 | 0,4867907 | 0,4871263 | 0,4874546 | 0,4877756 | 0,4880894 | 0,4883962 | 0,4886962 | 0,4889894 |
| 2,3 | 0,4892759 | 0,4895559 | 0,4898296 | 0,4900969 | 0,4903582 | 0,4906133 | 0,4908625 | 0,4911060 | 0,4913437 | 0,4915758 |
| 2,4 | 0,4918025 | 0,4920237 | 0,4922397 | 0,4924506 | 0,4926564 | 0,4928572 | 0,4930531 | 0,4932443 | 0,4934309 | 0,4936128 |
| 2,5 | 0,4937903 | 0,4939634 | 0,4941322 | 0,4942969 | 0,4944574 | 0,4946138 | 0,4947664 | 0,4949150 | 0,4950600 | 0,4952012 |
| 2,6 | 0,4953388 | 0,4954729 | 0,4956035 | 0,4957307 | 0,4958547 | 0,4959754 | 0,4960929 | 0,4962074 | 0,4963188 | 0,4964274 |
| 2,7 | 0,4965330 | 0,4966358 | 0,4967359 | 0,4968332 | 0,4969280 | 0,4970202 | 0,4971099 | 0,4971971 | 0,4972820 | 0,4973645 |
| 2,8 | 0,4974448 | 0,4975229 | 0,4975988 | 0,4976725 | 0,4977443 | 0,4978140 | 0,4978817 | 0,4979476 | 0,4980116 | 0,4980737 |
| 2,9 | 0,4981341 | 0,4981928 | 0,4982498 | 0,4983051 | 0,4983589 | 0,4984111 | 0,4984617 | 0,4985109 | 0,4985587 | 0,4986050 |
| 3,0 | 0,4986500 | 0,4986937 | 0,4987361 | 0,4987772 | 0,4988170 | 0,4988557 | 0,4988932 | 0,4989296 | 0,4989649 | 0,4989991 |
| 3,1 | 0,4990323 | 0,4990645 | 0,4990957 | 0,4991259 | 0,4991552 | 0,4991836 | 0,4992111 | 0,4992377 | 0,4992636 | 0,4992886 |
| 3,2 | 0,4993128 | 0,4993363 | 0,4993590 | 0,4993810 | 0,4994023 | 0,4994229 | 0,4994429 | 0,4994622 | 0,4994809 | 0,4994990 |
| 3,3 | 0,4995165 | 0,4995335 | 0,4995499 | 0,4995657 | 0,4995811 | 0,4995959 | 0,4996102 | 0,4996241 | 0,4996375 | 0,4996505 |
| 3,4 | 0,4996630 | 0,4996751 | 0,4996868 | 0,4996982 | 0,4997091 | 0,4997197 | 0,4997299 | 0,4997397 | 0,4997492 | 0,4997584 |
| 3,5 | 0,4997673 | 0,4997759 | 0,4997842 | 0,4997922 | 0,4997999 | 0,4998073 | 0,4998145 | 0,4998215 | 0,4998282 | 0,4998346 |
| 3,6 | 0,4998409 | 0,4998469 | 0,4998527 | 0,4998583 | 0,4998636 | 0,4998688 | 0,4998739 | 0,4998787 | 0,4998834 | 0,4998878 |
| 3,7 | 0,4998922 | 0,4998963 | 0,4999004 | 0,4999042 | 0,4999080 | 0,4999116 | 0,4999150 | 0,4999184 | 0,4999216 | 0,4999247 |
| 3,8 | 0,4999276 | 0,4999305 | 0,4999333 | 0,4999359 | 0,4999385 | 0,4999409 | 0,4999433 | 0,4999456 | 0,4999478 | 0,4999499 |
| 3,9 | 0,4999519 | 0,4999538 | 0,4999557 | 0,4999575 | 0,4999592 | 0,4999609 | 0,4999625 | 0,4999640 | 0,4999655 | 0,4999669 |
| 4,0 | 0,4999683 | 0,4999696 | 0,4999709 | 0,4999721 | 0,4999733 | 0,4999744 | 0,4999755 | 0,4999765 | 0,4999775 | 0,4999784 |

Em uma pesquisa de opinião, 32 dentre 80 homens declararam apreciar certa revista, acontecendo o mesmo com 26 dentre 50 mulheres. Ao nível de 5% de significância os homens e as mulheres apreciam igualmente a revista?

Solução:

As hipóteses são:

$$H_0: \pi_1 - \pi_2 = 0 \ (\pi_1 = \pi_2) \text{ contra}$$

$$H_1: \pi_1 - \pi_2 \neq 0 \ (\pi_1 \neq \pi_2)$$

$$\text{Tem-se que } P_1 = 32 / 80 = 0,40 \text{ e } P_2 = 26 / 50 = 52\%$$

O valor da variável teste será:

$$z = \frac{0,40 - 0,52}{\sqrt{\frac{0,40 \cdot 0,60}{80} + \frac{0,52 \cdot 0,48}{50}}} = -1,34$$

Como $\alpha = 5\%$, então $z\alpha/2 = -1,96$.

Portanto, aceita-se a hipótese de igualdade entre as preferências de homens e mulheres, isto é, a este nível de significância não é possível afirmar que exista diferença entre as preferências de homens e mulheres quanto à revista.

$$v = \frac{\left(\frac{S_X^2}{n} + \frac{S_Y^2}{m}\right)^2}{\frac{\left(\frac{S_X^2}{n}\right)^2}{n-1} + \frac{\left(\frac{S_Y^2}{m}\right)^2}{m-1}} = \frac{\left(\frac{7,5}{5} + \frac{5}{5}\right)^2}{\frac{\left(\frac{7,5}{5}\right)^2}{4} + \frac{\left(\frac{5}{5}\right)^2}{4}} = \frac{6,25}{0,8125} = 7,69 \cong 8,$$

então o valor de “ t ” tabelado será: 1,86.

Neste caso, com estas amostras não é possível afirmar que o concreto do tipo A seja mais resistente do que o concreto do tipo B.

Teste para a diferença entre duas proporções

As hipóteses são:

$H_0: \pi_1 - \pi_2 = \pi$ contra

$H_1: \pi_1 - \pi_2 \neq \pi$ ou

$\pi_1 - \pi_2 > \pi$ ou ainda

$\pi_1 - \pi_2 < \pi$

Se $\pi = 0$, então $\pi_1 - \pi_2 = 0$, isto é, $\pi_1 = \pi_2$.

Extraídas uma amostra de cada uma das duas populações a variável $P_1 - P_2$ terá uma distribuição aproximadamente normal com média $E(P_1 - P_2) = \pi_1 - \pi_2$ e variância $\sigma_{P_1 - P_2}^2$

$$= \frac{\pi_1(1-\pi_1)}{n} + \frac{\pi_2(1-\pi_2)}{m}, \text{ desde que } nP_1 > 5 \text{ e } mP_2 > 5.$$

A variável teste será, então:
$$z = \frac{P_1 - P_2 - \pi}{\sqrt{\frac{\pi_1(1-\pi_1)}{n} + \frac{\pi_2(1-\pi_2)}{m}}}$$

Como os valores de π_1 e π_2 não são conhecidos, deve-se utilizar suas estimativas P_1 e P_2 . Desta forma, o valor de z será:

$$z = \frac{P_1 - P_2 - \pi}{\sqrt{\frac{P_1(1-P_1)}{n} + \frac{P_2(1-P_2)}{m}}}$$

Assim fixando o nível de significância “ α ”, a hipótese nula será rejeitada se:

$|z| > z_{\alpha/2}$ no teste bilateral;

$z > z_{\alpha}$, no teste unilateral à direita e

$z < z_{\alpha}$ no teste unilateral à esquerda.

Exemplo:

$$v = \frac{\left(\frac{S_X^2}{n} + \frac{S_Y^2}{m} \right)^2}{\frac{\left(\frac{S_X^2}{n} \right)^2}{n-1} + \frac{\left(\frac{S_Y^2}{m} \right)^2}{m-1}}$$

Desde que n, m sejam maiores ou iguais a 30, ou então que as amostras tenham sido extraídas de populações que tenham distribuições normais.

Assim fixando o nível de significância “ α ”, a hipótese nula será rejeitada se:

$|t_c| > t_{\alpha/2}$ no teste bilateral;

$t_c > t_\alpha$, no teste unilateral à direita e

$t < t_\alpha$ no teste unilateral à esquerda, onde $t = \frac{\bar{X} - \bar{Y} - \Delta}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}}$

Exemplo:

As resistências de dois tipos de concreto foram medidas, mostrando os resultados da tabela. Fixado um nível de significância de 5%, existe evidências de que o concreto do tipo A seja mais resistente do que o concreto do tipo B.

| | | | | | |
|---------------|----|----|----|----|----|
| Tipo A | 54 | 55 | 58 | 51 | 57 |
| Tipo B | 50 | 54 | 56 | 52 | 53 |

Solução:

As hipóteses são:

$H_0: \mu_A - \mu_B = 0$ ($\mu_A = \mu_B$) contra

$H_1: \mu_A - \mu_B > 0$ ($\mu_A > \mu_B$)

Os dados obtidos da tabela são:

$\bar{X} = 55,0$ e $\bar{Y} = 53,0$

$S_X^2 = 7,50$ e $S_Y^2 = 5,0$

O valor da variável teste será:

$$t = \frac{55 - 53}{\sqrt{\frac{7,5}{5} + \frac{5}{5}}} = 1,265$$

Com $\alpha = 5\%$, e o grau de liberdade

Solução:

As hipóteses são:

$$H_0: \mu_A - \mu_B = 0 \quad (\mu_A = \mu_B) \text{ contra}$$

$$H_1: \mu_A - \mu_B > 0 \quad (\mu_A > \mu_B)$$

Os dados obtidos da tabela são:

$$\bar{X} = 55,0 \text{ e } \bar{Y} = 53,0$$

$$S_X^2 = 7,50 \text{ e } S_Y^2 = 5,0, \text{ então } S^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2} = \frac{(5-1).7,5 + (5-1).5,0}{5+5-2} = 6,25.$$

O valor da variável teste será:

$$t_c = \frac{55 - 53}{2,50 \cdot \sqrt{\frac{1}{5} + \frac{1}{5}}} = 1,265$$

Como $\alpha = 5\%$, e o grau de liberdade $n - m - 2 = 10 - 2 = 8$, então o valor de “t” tabelado será: 1,86.

Neste caso, com estas amostras não é possível afirmar que o concreto do tipo A seja mais resistente do que o concreto do tipo B.

(c) Variâncias populacionais σ_X^2 e σ_Y^2 desconhecidas e supostamente desiguais.

As hipóteses são:

$$H_0: \mu_X - \mu_Y = \Delta \text{ contra}$$

$$H_1: \mu_X - \mu_Y \neq \Delta \text{ ou}$$

$$\mu_X - \mu_Y > \Delta \text{ ou ainda}$$

$$\mu_X - \mu_Y < \Delta$$

Como as variâncias são desconhecidas é necessária estimá-las através das variâncias amostrais S_X^2 e S_Y^2 . Neste caso, ao se substituir as variâncias populacionais pelas amostrais na expressão:

$$\frac{\bar{X} - \bar{Y} - \Delta}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}$$

Não se terá mais uma distribuição normal, mas sim uma distribuição “t” com o grau de liberdade fornecido pela seguinte expressão:

$$H_1: \mu_X - \mu_Y \neq \Delta \text{ ou}$$

$$\mu_X - \mu_Y > \Delta \text{ ou ainda}$$

$$\mu_X - \mu_Y < \Delta$$

A variável teste anterior, para esta situação, será:

$$Z = \frac{\bar{X} - \bar{Y} - \Delta}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}, \text{ mas neste caso } \sigma_X^2 = \sigma_Y^2 = \sigma^2 \text{ (por suposição), então:}$$

$$Z = \frac{\bar{X} - \bar{Y} - \Delta}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} = \frac{\bar{X} - \bar{Y} - \Delta}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}} = \frac{\bar{X} - \bar{Y} - \Delta}{\sigma \sqrt{\frac{1}{n} + \frac{1}{m}}}, \text{ como o valor } \sigma^2 \text{ não é conhecido, deverá ser}$$

substituído por um estimador não-tendencioso. Como S_X^2 e S_Y^2 são estimadores não tendenciosos do mesmo parâmetro σ^2 , então, a média ponderada:

$$S^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}, \text{ também será um estimador não-tendencioso de } \sigma^2.$$

Logo a expressão acima poderá ser escrita como:

$$\frac{\bar{X} - \bar{Y} - \Delta}{S \sqrt{\frac{1}{n} + \frac{1}{m}}}, \text{ que terá uma distribuição não mais normal mas sim “t” com “n + m - 2”}$$

graus de liberdade, desde que n, m sejam maiores ou iguais a 30, ou então que as amostras tenham sido extraídas de populações que tenham distribuições normais.

Desta forma, a expressão para testar a diferença entre duas médias populacionais, nesta situação será:

$$t_c = t_{n+m-2} = \frac{\bar{X} - \bar{Y} - \Delta}{S \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

Assim fixando o nível de significância “ α “, a hipótese nula será rejeitada se:

$$|t_c| > t_{\alpha/2} \text{ no teste bilateral;}$$

$$t_c > t_{\alpha}, \text{ no teste unilateral à direita e}$$

$$t_c < t_{\alpha} \text{ no teste unilateral à esquerda.}$$

Exemplo:

As resistências de dois tipos de concreto foram medidas, mostrando os resultados da tabela. Fixado um nível de significância de 5%, existe evidência de que o concreto do tipo A seja mais resistente do que o concreto do tipo B.

| | | | | | |
|---------------|----|----|----|----|----|
| Tipo A | 54 | 55 | 58 | 51 | 57 |
| Tipo B | 50 | 54 | 56 | 52 | 53 |

Como as variâncias são conhecidas, tem-se então que, para $n, m \geq 30$ ou para amostras extraídas de populações normais, que a variável $\bar{D} = \bar{X} - \bar{Y}$ terá uma distribuição

aproximadamente normal com média $E(\bar{D}) = \mu_X - \mu_Y$ e variância $V(\bar{D}) = \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}$.

A variável teste será, então:

$$z = \frac{\bar{X} - \bar{Y} - \Delta}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}$$

Assim fixando o nível de significância “ α ”, a hipótese nula será rejeitada se:

$|z| > z_{\alpha/2}$ no teste bilateral;

$z > z_{\alpha}$, no teste unilateral à direita e

$z < z_{\alpha}$ no teste unilateral à esquerda.

Exemplo:

Um fabricante produz dois tipos de pneus. Para o pneu do tipo A o desvio padrão é de 2500 km e para o pneu do tipo B é de 3000 km. Uma cia de táxis testou 50 pneus do tipo A e 40 do tipo B, obtendo 24000 km de média para o “A” e 26000 para o tipo “B”.

Adotando $\alpha = 4\%$ testar a hipótese de que a duração média dos dois tipos é a mesma.

Solução:

As hipóteses são:

$H_0: \mu_A - \mu_B = 0$ ($\mu_A = \mu_B$) contra

$H_1: \mu_A - \mu_B \neq 0$ ($\mu_A \neq \mu_B$)

Como $\alpha = 4\%$, então $z_{\alpha/2} = 2,05$.

O valor da variável teste será:

$$z = \frac{24000 - 26000}{\sqrt{\frac{2500^2}{50} + \frac{3000^2}{40}}} = -3,38$$

Portanto, rejeita-se a hipótese de igualdade entre as durações médias dos dois tipos de pneus. Com base nestas amostras, pode-se afirmar, ao nível de 4% de significância, que os dois tipos de pneus diferem quanto a durabilidade média.

(b) Variâncias populacionais σ_X^2 e σ_Y^2 desconhecidas mas supostamente iguais

Vamos supor que as duas populações tenham a mesma variância $\sigma^2 = \sigma_X^2 = \sigma_Y^2$, porém desconhecidas.

As hipóteses são:

$H_0: \mu_X - \mu_Y = \Delta$ contra

Exemplo

Uma das maneiras de controlar a qualidade de um produto é controlar a sua variabilidade. Uma máquina de empacotar café está regulada para encher os pacotes com desvio padrão de 10 g e média de 500g e onde o peso de cada pacote distribuí-se normalmente. Colhida uma amostra de $n = 16$, observou-se uma variância de 169 g^2 . É possível afirmar com este resultado que a máquina está desregulada quanto a variabilidade, supondo uma significância de 5%?

Solução

$$H_0: \sigma^2 = 100 \text{ contra}$$

$$H_1: \sigma^2 \neq 100$$

$$\chi_c^2 = (15.169)/100 = 25,35.$$

Como $\alpha = 5\%$ a região de aceitação é a região compreendida entre os valores: $[\chi_{97,5\%}^2, \chi_{2,5\%}^2] = [6,26, 27,49]$. Como o valor calculado pertence a esta região, aceita-se H_0 , isto é, com esta amostra não é possível afirmar que a máquina está desregulada, ao nível de 5% de significância.

Supõem-se a existência de duas populações. Uma população X com média μ_X e desvio padrão σ_X e uma população Y com média μ_Y e desvio padrão σ_Y . Da população X é extraída uma amostra de tamanho “n” com média \bar{X} e da população Y é extraída uma amostra de tamanho “m” com média \bar{Y} . Define-se a variável \bar{D} como sendo a diferença entre as duas médias amostrais. Assim $\bar{D} = \bar{X} - \bar{Y}$ e tem-se:

$$\mu_{\bar{D}} = E(\bar{D}) = E(\bar{X} - \bar{Y}) = E(\bar{X}) - E(\bar{Y}) = \mu_X - \mu_Y$$

$$\sigma_{\bar{D}} = V(\bar{D}) = V(\bar{X} - \bar{Y}) = V(\bar{X}) + V(\bar{Y}) = \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}$$

Teste para a diferença entre duas médias

(a) Conhecidas as variâncias populacionais σ_X^2 e σ_Y^2

As hipóteses são:

$$H_0: \mu_X - \mu_Y = \Delta \text{ contra}$$

$$H_1: \mu_X - \mu_Y \neq \Delta \text{ ou}$$

$$\mu_X - \mu_Y > \Delta \text{ ou ainda}$$

$$\mu_X - \mu_Y < \Delta$$

Se $\Delta = 0$, então $\mu_X - \mu_Y = 0$, isto é, $\mu_X = \mu_Y$.

Considerando, então, um teste bilateral e tendo $\alpha = 5\%$ tem-se que a região de aceitação é constituída pelo intervalo $RA = [-1,96, 196]$.

O valor de teste é:

$$z = \frac{p - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} = \frac{0,53 - 0,60}{\sqrt{\frac{0,60(1-0,60)}{1000}}} = -4,42.$$

Como este valor não pertence a região de aceitação, pode-se rejeitar a hipótese nula, ao nível de 5% de significância, isto é, neste caso, pode-se afirmar que a taxa dos que sobrevivem até os 60 anos é menor do que 60%. Neste caso, também poderia ser realizado um teste unilateral à esquerda. Este teste também rejeitaria a hipótese nula, pois para ele o valor crítico $z_{\alpha} = -1,645$.

Teste para a variância de uma população

Para aplicar o teste para a variância é necessário supor a normalidade da população de onde será extraída a amostra.

As hipóteses são:

$$H_0: \sigma^2 = \sigma_0^2 \text{ contra}$$

$$H_1: \sigma^2 \neq \sigma_0^2$$

$$\sigma^2 > \sigma_0^2$$

$$\sigma^2 < \sigma_0^2$$

$$\text{A estatística teste é } \frac{(n-1)S^2}{\sigma_0^2} \sim \chi_{n-1}^2$$

Quer dizer o quociente acima tem uma distribuição qui-quadrado com “n-1” graus de liberdade. A qui-quadrado é uma distribuição assimétrica positiva que varia de zero a mais infinito. Esta distribuição é tabelada também em função dos número de graus de liberdade, isto é, cada grau de liberdade (n - 1) representa uma distribuição diferente. As colunas das tabelas representam diferentes níveis de significância, isto é, área sob a curva acima do valor tabelado.

Em função do tipo de hipótese alternativa define-se a região de rejeição. No primeiro caso tem-se uma região de rejeição do tipo bilateral. Logo, fixado um nível de significância “ α “, a região crítica será $RC = [0, \chi_1^2] \cup [\chi_2^2, \infty)$. Desta forma, aceita-se a hipótese nula se a estatística teste, acima, pertencer ao intervalo $[\chi_1^2, \chi_2^2]$.

Considerando, então, um teste unilateral à esquerda e tendo $\alpha = 5\%$ ($\alpha = 1\%$) tem-se que a região de rejeição é constituída por $RC = [-\infty, -1,753]$. ($RC = [-\infty, -2,602]$)

O valor de teste é:

$$t_{15} = \frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{85 - 100}{12/\sqrt{4}} = -5$$

Como este valor pertence as duas regiões críticas, pode-se rejeitar a hipótese nula, aos níveis de 5% e 1% de significância, isto é, neste caso, pode-se afirmar que a modificação diminuiu o tempo de execução da tarefa.

Teste para a proporção

O teste para a proporção populacional é normalmente baseado na seguinte suposição: tem-se uma população e tem-se uma hipótese sobre a proporção π de elementos da população que possuem uma determinada característica. Esta proporção é supostamente igual a um determinado valor π_0 . Assim a hipótese nula é:

$$H_0 : \pi = \pi_0$$

O problema fornece informações sobre a alternativa, que pode ser uma das seguintes:

$$H_1 : \pi \neq \pi_0$$

$$H_1 : \pi > \pi_0$$

$$H_1 : \pi < \pi_0$$

A estatística teste a ser utilizada é a proporção amostral “P”, que para amostras grandes ($n > 50$) tem uma distribuição aproximadamente normal com média:

$\mu_P = \pi$, e desvio padrão

$$\sigma_P = \sqrt{\frac{\pi(1-\pi)}{n}}$$

Exemplo:

As condições de mortalidade de uma região são tais que a proporção de nascidos que sobrevivem até 60 anos é de 0,60. Testar esta hipótese ao nível de 5% de significância se em 1000 nascimentos amostrados aleatoriamente, verificou-se 530 sobreviventes até os 60 anos.

Solução

$$H_1: \pi = 0,60$$

$$H_0: \pi \neq 0,60$$

De fato, conforme demonstrado por W. S. Gosset (Student) a distribuição da variável:

$$(\bar{X} - \mu_{\bar{X}}) / \hat{\sigma}_{\bar{X}} = (\bar{X} - \mu) / s / \sqrt{n}$$

Não é mais normal padrão. Ao substituir σ por s na expressão teremos uma distribuição parecida com a normal, isto é, simétrica em torno de zero, porém com uma variabilidade maior. Desta forma a distribuição “ t ” é mais baixa no centro do que a normal padrão, mas mais alta nas caudas.

Assim:

$$(\bar{X} - \mu_{\bar{X}}) / \hat{\sigma}_{\bar{X}} = (\bar{X} - \mu) / s / \sqrt{n} = t_{n-1}, \text{ onde “} n - 1 \text{” indica a distribuição “} t \text{”}$$

considerada, pois cada tamanho de amostra produz uma distribuição de Student diferente.

A distribuição t de Student encontra-se tabelada em função de n = tamanho da amostra ou então em função de $n - 1$ denominado de graus de liberdade da distribuição. Neste caso cada linha de uma tabela se refere a uma distribuição particular e cada coluna da tabela a um determinado nível de significância. Conforme a tabela o nível de significância poderá ser unilateral ou bilateral. Em todo caso é necessário sempre ler no cabeçalho ou no rodapé da tabela as explicações sobre como ela está estruturada.

Desta forma a diferença entre o teste para a média de uma população com σ conhecido e um com σ desconhecido é que é necessário trocar a distribuição normal padrão pela distribuição “ t ” de Student.

Exemplo:

O tempo médio, por operário, para executar uma tarefa, tem sido 100 minutos. Introduziu-se uma modificação para diminuir este tempo, e, após certo período, sorteu-se uma amostra de 16 operários, medindo-se o tempo de execução gasto por cada um. O tempo médio da amostra foi 85 minutos com desvio padrão de 12 minutos. Este resultado evidencia uma melhora no tempo gasto para realizar a tarefa? Apresente as conclusões aos níveis de 5% e 1% de significância e diga quais as suposições teóricas necessárias que devem ser feitas para resolver o problema.

Solução

A suposição teórica necessária é admitir que a distribuição da população de onde foi extraída a amostra segue uma normal pois $n < 30$.

$$H_0: \mu = 100$$

$$H_1: \mu < 100$$

programa de prevenção de acidentes e, após o mesmo, tomou-se uma amostra de 9 indústrias e mediu-se o número de horas/homem perdidas por acidente, que foi de 50 horas. Você diria, ao nível de 5%, que há evidência de melhoria?

Solução

As hipóteses a serem testadas são:

$$H_0: \mu = 60 \text{ hora/homens}$$

$$H_1: \mu < 60 \text{ hora/homens}$$

A evidência amostral para sugerir que a média baixou é dada através da amostra de $n = 9$ (elementos) que forneceu $\bar{X} = 50$ horas/homens. Vamos testar se esta diferença de 10 horas/homens é ou não significativa ao nível de 5%. Para isto é necessário padronizar o resultado amostral.

$$Z = (\bar{X} - \mu_{\bar{X}}) / \sigma_{\bar{X}} = (\bar{X} - \mu) / \sigma / \sqrt{n} = (50 - 60) / 20 / \sqrt{9} = -1,50$$

Para saber se este valor (-1,50) é pouco provável é necessário compará-lo com o valor crítico - z_{α} (pois se trata de um teste unilateral à esquerda), que neste caso vale -1,64, já que o nível de significância foi fixado em 5%. Vê-se portanto que o valor amostral não é inferior ao valor crítico, não estando portanto na região de rejeição. Isto quer dizer que a diferença apresentada na amostra não é suficientemente grande para provar que a campanha de prevenção deu resultado. Então a conclusão é:

“Não é possível ao nível de 5% de significância afirmar que a campanha deu resultado, isto é, rejeitar H_0 . ”

Convém lembrar que o fato de não rejeitar a hipótese nula, não autoriza a fazer afirmações a respeito da veracidade dela. Ou seja, não se provou H_0 , pois no momento que se aceita a hipótese nula, o risco envolvido é o do Tipo II, e este neste caso não está fixado (controlado). O teste de hipóteses é feito para rejeitar a hipótese nula e sua força está na rejeição. Assim quando se rejeita se prova algo, mas quando se aceita, nada se pode afirmar.

(b) σ desconhecido

A distribuição t de Student

Quando o desvio padrão populacional (σ) é desconhecido é necessário estimá-lo através do desvio padrão da amostra (s). Mas ao substituir o desvio padrão da população na expressão:

$$Z = (\bar{X} - \mu_{\bar{X}}) / \sigma_{\bar{X}} = (\bar{X} - \mu) / \sigma / \sqrt{n}$$

não teremos mais uma distribuição normal.

$$\mu < \mu_0$$

$$\mu \neq \mu_0$$

A estatística teste utilizada aqui é a média da amostra: \bar{X} . Esta média para ser comparada com o valor tabelado, determinado em função da probabilidade do erro do tipo I, (isto é, o nível de significância do teste), precisa ser primeiramente padronizada. Isto é feito, baseado no seguinte resultado:

Se X é uma variável aleatória normal com média μ e desvio padrão σ , então a variável:

$$Z = (X - \mu) / \sigma$$

Tem uma distribuição normal com média “0” e desvio padrão “1”. A variável resultante Z se encontra tabelada. Qualquer livro de Estatística traz esta tabela que fornece os valores desta variável, para z variando de -3,9 até 3,9 em intervalos de 0,1 (aproximação decimal), entre -3,9 e -3,0 e entre 3,0 e 3,9, e em intervalos de 0,01 (aproximação centesimal) para os valores entre -3,0 e 3,0.

Para \bar{X} sabe-se que $\mu_{\bar{X}} = \mu$ (média das médias) que $\sigma_{\bar{X}} = \sigma/\sqrt{n}$ (erro padrão da média), então o valor padronizado de \bar{X} será:

$$Z = (\bar{X} - \mu_{\bar{X}}) / \sigma_{\bar{X}} = (\bar{X} - \mu) / \sigma/\sqrt{n}$$

Supondo-se fixado um nível de significância de $\alpha = P(\text{Erro do Tipo I})$, verifica-se na tabela qual o valor de z_{α} (no teste unilateral) ou $z_{\alpha/2}$ (teste bilateral). Rejeita-se H_0 (hipótese nula) se o valor de z calculado na expressão acima for:

- (i) Maior do que z_{α} (no teste unilateral à direita);
- (ii) Menor do $-z_{\alpha}$ (no teste unilateral à esquerda) e
- (iii) Maior que $z_{\alpha/2}$ ou menor que $-z_{\alpha/2}$ (no teste bilateral).

Tabela 03 - Valores de z para alguns níveis de significância

| | $\alpha = \text{Nível de significância} = P(\text{Erro do Tipo I})$ | | |
|-------------------------|---|------|------|
| | 10% | 5% | 1% |
| Teste bilateral | 1,64 | 1,96 | 2,57 |
| Teste unilateral | 1,28 | 1,64 | 2,33 |

Exemplo:

A associação dos proprietários de indústrias metalúrgicas está preocupada com o tempo perdido em acidentes de trabalho, cuja média, nos últimos tempos, tem sido da ordem de 60 hora /homens por ano com desvio padrão de 20 horas/homem. Tentou-se um

Após fixar as hipóteses é necessário determinar se a diferença entre a estatística amostral e o suposto valor do parâmetro da população é suficiente para rejeitar a hipótese. A estatística utilizada deve ser definida e sua distribuição teórica determinada.

3. Fixar o nível de significância do teste.

Fixar a probabilidade de ser cometer erro do tipo I, isto é, estabelecer o nível de significância do teste. Fixado o erro do tipo I, é possível determinar o valor crítico, que é um valor lido na distribuição amostral da estatística considerada (tabela). Este valor vai separar a região de crítica (de rejeição) da região de aceitação.

4. Calcular a estatística teste (a estimativa).

Através da amostra obtida calcular a estimativa que servirá para aceitar ou rejeitar a hipótese nula. Dependendo do tipo de hipótese alternativa este valor servirá para aceitar ou rejeitar H_0 . O procedimento é:

| |
|--|
| $\text{Teste estatístico} = (\text{Estatística} - \text{Parâmetro}) / \text{Erro padrão da Estatística}$ |
|--|

5. Tomar a decisão.

Se o valor da estatística estiver na região crítica rejeitar H_0 , caso contrário, aceitar H_0 .

TIPOS DE TESTES PARAMÉTRICOS

Os testes paramétricos podem ser divididos em testes para:

- Uma amostra
- Duas amostras emparelhadas (dependentes)
- Duas amostras independentes
- Várias amostras (Análise de Variância)

Testes para uma amostra

Teste para a média de uma população

(a) σ conhecido

O teste para a média de uma população pode ser executado com qualquer tamanho de amostra se soubermos que a população de onde for extraída a amostra segue uma distribuição normal. Se a distribuição da população não for conhecida então é necessário trabalhar com amostras grandes (pelo menos 30 elementos) para poder garantir a normalidade da média da amostra através do teorema central do limite.

As hipóteses são:

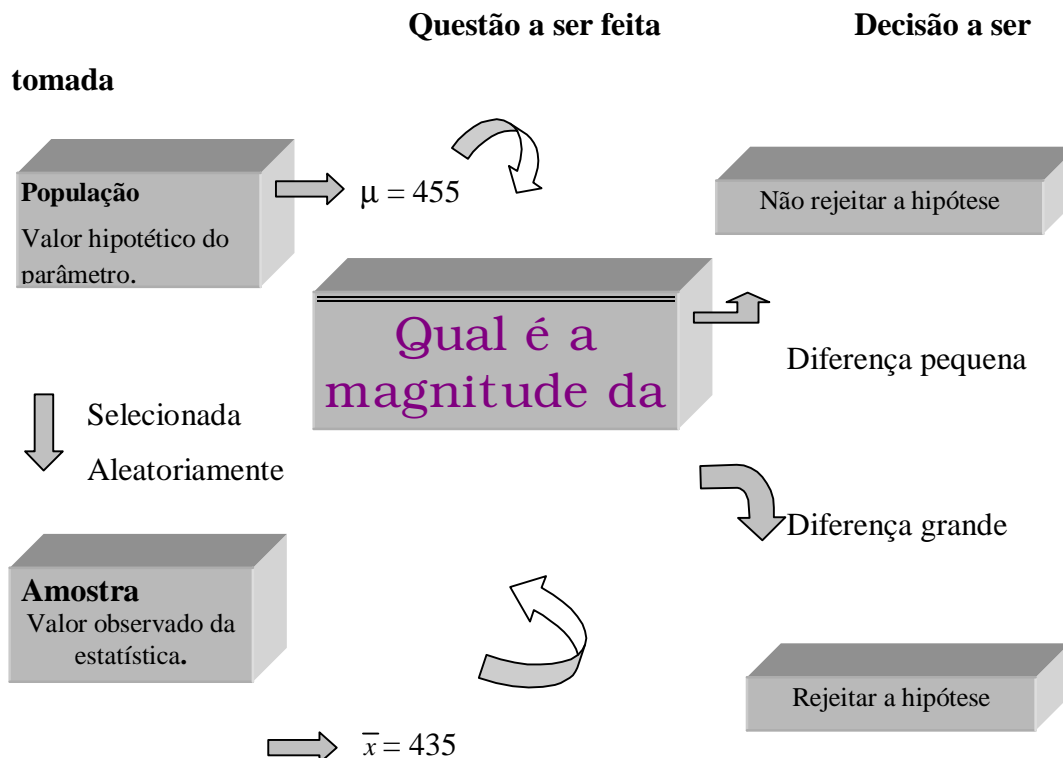
$H_0: \mu = \mu_0$ contra

$H_1: \mu = \mu_1$ ou então, o que é mais comum:

$H_1: \mu > \mu_0$

Etapas do teste de hipóteses

Qualquer teste de hipóteses paramétrico segue os seguintes passos:



1. Formular as hipóteses.

Estabelecer as hipóteses nula e alternativa. A construção de um teste de hipóteses pode ser colocado de forma geral do seguinte modo. Toma-se uma amostra da variável (ou das variáveis) X (no caso) de uma dada população, de onde se tem uma hipótese sobre um determinado parâmetro, por exemplo: θ . Esta hipótese é a hipótese nula ou hipótese de igualdade:

$$H_0: \theta = \theta_0$$

Tendo formulado a hipótese nula é conveniente determinar qual será a hipótese aceita caso a hipótese nula seja rejeitada, isto é, convém explicitar a hipótese alternativa. A hipótese alternativa vai depender de cada situação mas de forma geral tem-se:

$H_1: \theta = \theta_2$ (hipótese simples), ou então o que é mais comum, hipóteses compostas:

$H_1: \theta > \theta_0$ (teste unilateral ou unicaudal à direita)

$\theta < \theta_0$ (teste unilateral ou unicaudal à esquerda)

$\theta \neq \theta_0$ (teste bilateral ou bicaudal) as hipóteses são do tipo composto.

2. Estabelecer a estatística (estimador) a ser utilizado.

infinitos possíveis. É natural que os erros associados sejam grandes, pois a amostra é muito pequena. Aumentado-se o tamanho da amostra é possível com a mesma região crítica diminuir sensivelmente os dois tipos de erro.

A distribuição amostral

A distribuição amostral é uma distribuição de probabilidade, isto é, é uma distribuição teórica que descreve o comportamento de uma determinada estatística ou estimador. As principais estatísticas utilizadas nos testes de hipóteses possuem modelos conhecidos. Têm-se a distribuição normal, a distribuição t (de Student) a distribuição χ^2 (qui-quadrado), a distribuição F (de Snedkor) como as principais.

Testes estatísticos paramétricos

Em termos gerais, uma hipótese é uma conjectura sobre algum fenômeno ou conjunto de fatos. Em estatística inferencial o termo *hipótese* tem um significado bastante específico. É uma conjectura sobre uma ou mais parâmetros populacionais. O teste de hipóteses paramétrico envolve fazer inferências sobre a natureza da população com base nas observações de uma amostra extraída desta população.

Figura 11- A lógica do teste de hipóteses

Em outras palavras, testar hipóteses, envolve determinar a magnitude da diferença entre um valor observado de uma estatística, por exemplo a proporção p , e o suposto valor do parâmetro (π) e então decidir se a magnitude da diferença justifica a rejeição da hipótese. O processo segue o esquema da figura 01.

$$\begin{aligned}
 1 - \alpha &= P(\text{Decisão correta}) = P(\text{Aceitar } H_0 / H_0 \text{ é verdadeira}) \\
 &= P(x \in RA / p = 50\%) = P(x \in \{0, 1, 2\} / p = 50\%) \\
 &= 1/32 + 5/32 + 10/32 = 16/32 = 50\%
 \end{aligned}$$

$$\begin{aligned}
 \beta &= P(\text{Erro do tipo II}) = P(\text{Aceitar } H_0 / H_0 \text{ é falsa}) \\
 &= P(x \in RA / p = 80\%) = P(x \in \{0, 1, 2\} / p = 80\%) \\
 &= 1/3125 + 20/3125 + 160/3125 = 181/3125 = 5,69\%
 \end{aligned}$$

$$\begin{aligned}
 1 - \beta &= P(\text{Erro do tipo II}) = P(\text{Rejeitar } H_0 / H_0 \text{ é falsa}) \\
 &= P(x \in RC / p = 80\%) = P(x \in \{3, 4, 5\} / p = 80\%) \\
 &= 640/3125 + 1280/3125 + 1024/3125 = 2944/3125 = 94,31\%
 \end{aligned}$$

Por estes resultados pode-se verificar, que o erro do tipo II poderia ser aceitável, mas o erro do tipo I não, pois é um valor igual a probabilidade de se decidir corretamente.

Neste caso, uma opção para diminuir o erro do tipo I seria mudar a região de rejeição.

Se a região crítica escolhida tivesse sido $RC = \{5\}$, isto é, rejeitar a hipótese nula somente se em 5 lançamentos da moeda fosse obtida 5 caras as probabilidades acima ficariam:

$$\begin{aligned}
 \alpha &= \text{nível de significância do teste} = P(\text{Erro do tipo I}) \\
 &= P(\text{rejeitar } H_0 / H_0 \text{ é verdadeira}) = P(x \in RC / p = 50\%) \\
 &= P(x \in \{5\} / p = 50\%) = 1/32 = 3,12\%
 \end{aligned}$$

$$\begin{aligned}
 1 - \alpha &= P(\text{Decisão correta}) = P(\text{Aceitar } H_0 / H_0 \text{ é verdadeira}) \\
 &= P(x \in RA / p = 50\%) = P(x \in \{0, 1, 2, 3, 4\} / p = 50\%) \\
 &= 1/32 + 5/32 + 10/32 + 10/32 + 5/32 = 31/32 = 96,88\%
 \end{aligned}$$

$$\begin{aligned}
 \beta &= P(\text{Erro do tipo II}) = P(\text{Aceitar } H_0 / H_0 \text{ é falsa}) \\
 &= P(x \in RA / p = 80\%) = P(x \in \{0, 1, 2, 3, 4\} / p = 80\%) \\
 &= 1/3125 + 20/3125 + 160/3125 + 640/3125 + 1280/3125 = \\
 &= 2101/3125 = 67,33\%
 \end{aligned}$$

$$\begin{aligned}
 1 - \beta &= P(\text{Erro do tipo II}) = P(\text{Rejeitar } H_0 / H_0 \text{ é falsa}) \\
 &= P(x \in RC / p = 80\%) = P(x \in \{5\} / p = 80\%) \\
 &= 1024/3125 = 32,77\% = \text{Poder do teste.}
 \end{aligned}$$

Pode-se ver então que o erro do tipo I diminui sensivelmente, mas em compensação tivemos um aumento substancial do erro do tipo II. Isto sempre vai ocorrer. A única forma de reduzir os dois tipos de erro simultaneamente é pelo aumento do tamanho da amostra. Neste caso, está se considerando uma amostra de apenas 5 lançamentos dos

A probabilidade de que a variável (número de caras) assuma um valor do conjunto RC é denominada de **nível de significância do teste**. O nível de significância do teste é, na realidade, a probabilidade de se rejeitar a hipótese nula, quando ela é verdadeira, sendo então a probabilidade de se cometer um erro. Como este é apenas um dos dois tipos de erro possível de ser cometido num teste de hipóteses, ele é denominado de **erro do tipo I**. O outro tipo de erro possível de ser cometido é aceitar H_0 quando ela é falsa e é denominado de **erro do tipo II**. Em resumo pode-se ter as seguintes situações em um teste de hipóteses:

Tabela 2 - Possibilidades envolvidas em um teste de hipóteses

| | | Decisão | |
|-----------|--------------------|--|---|
| | | Aceitar H_0 | Rejeitar H_0 |
| Realidade | H_0 é verdadeira | Decisão correta ($1 - \alpha$) $= P(\text{Aceitar } H_0 / H_0 \text{ é verdadeira})$ $= P(H_0 / H_0)$ | Erro do Tipo I (α) $= P(\text{Cometer Erro do tipo I})$ $= P(\text{Rejeitar } H_0 / H_0 \text{ é verdadeira})$ $= \text{Nível de significância do teste}$ $= P(H_1 / H_0)$ |
| | H_0 é falsa | Erro do Tipo II $\beta = P(\text{Cometer Erro do tipo II})$ $= P(\text{Aceitar } H_0 / H_0 \text{ é falsa})$ $= P(\text{Aceitar } H_0 / H_1 \text{ é verdadeira})$ $= P(H_0 / H_1)$ | Decisão correta $1 - \beta = P(\text{Rejeitar } H_0 / H_0 \text{ é falsa})$ $= P(H_1 / H_1) = \text{Poder do teste.}$ |

Pode-se, agora, determinar as probabilidades de se cometer os erros dos tipos I e II e como consequência às probabilidades de se tomar às decisões corretas. A probabilidade de se cometer erro do tipo II, pode ser determinada aqui, porque o teste é do tipo simples, isto é, a hipótese alternativa envolve um único valor (neste caso $p = 80\%$). Geralmente, a hipótese alternativa é do tipo composto ($p < 80\%$ ou $p > 80\%$ ou ainda $p \neq 80\%$), e então a determinação do erro do tipo II só poderá ser feita mediante suposições à respeito dos valores que ela pode assumir. Existirão, na realidade, infinitas opções para o erro do tipo II. Para este caso, tem-se:

$\alpha = \text{nível de significância do teste} = P(\text{Erro do tipo I})$

$$= P(\text{rejeitar } H_0 / H_0 \text{ é verdadeira}) = P(x \in \text{RC} / p = 50\%)$$

$$= P(x \in \{3, 4, 5\} / p = 50\%) = 10/32 + 5/32 + 1/32 = 16/32 = 50\%$$

Para poder aceitar ou rejeitar H_0 e como consequência, rejeitar ou aceitar H_1 , é necessário estabelecer uma regra de decisão, isto é, é necessário estabelecer para que valores da variável X vai-se rejeitar H_0 , ou seja, afirmar H_1 , e para que valores da variável X , vai-se aceitar H_0 , ou seja, nesta situação particular, afirmar H_0 .

Desta forma, estabelecendo-se que se vai rejeitar H_0 , se a moeda lançada der um número de caras igual a 3, 4 ou 5, pode-se então determinar as probabilidades de tomar as decisões corretas ou as probabilidades dos erros envolvidos. Assim o conjunto de valores que levará a rejeição da hipótese nula será denominado de **região crítica (RC)** e, neste caso, este conjunto é igual a:

$$RC = \{ 3, 4, 5 \}$$

A faixa restante de valores da variável é denominada de **região de aceitação (RA)** e, neste caso, este conjunto vale:

$$RA = \{ 0, 1, 2 \}$$

Evidentemente esta regra como qualquer outra permitirá decidir sob a H_0 , mas estará sujeita a erro. Está se tomando a decisão de aceitar ou rejeitar H_0 com base no número X de caras obtidas em 5 lançamentos, que é apenas uma amostra, muito pequena, do número infinito de lançamentos possíveis.

Com base em resultados amostrais, não é possível tomar decisões definitivamente corretas. Entretanto, pode-se calcular a probabilidade da decisão estar errada. Neste caso foi decidido rejeitar H_0 se $X =$ “número de caras” assumir um dos valores do conjunto RC. No entanto, tais valores podem ocorrer sob H_0 , isto é, tais valores podem ocorrer quando se lança a moeda M_1 , conforme tabela. Então se H_0 for rejeitada porque X assumiu o valor 3, 4 ou 5, pode-se estar cometendo um erro. A probabilidade deste erro é igual a probabilidade de ocorrência destes valores sob H_0 , isto é, quando a moeda M_1 é lançada, que é conforme tabela igual a:

$$10/32 + 5/32 + 1/32 = 16/32 = 50\%$$

Lembrando que rejeitar H_0 é apenas uma das duas situações possíveis num teste de hipóteses, tem-se que se X assumir um valor do conjunto RA se aceitará H_0 . Mas tais valores podem ocorrer sob H_1 , isto é, quando a moeda M_2 é lançada. Então se H_0 for aceita porque X assumiu um dos valores: 1, 2 ou 3, pode-se estar cometendo um outro tipo de erro, cuja probabilidade é igual a da ocorrência destes valores sob H_1 que é de:

$$1/3125 + 20/3125 + 160/3125 = 181/3125 = 5,69\%$$

probabilidade de rejeitar H_0 quando for falsa, são as que exigem as suposições mais fortes ou mais amplas.

Conceitos adicionais do teste de hipóteses

Além dos conceitos já vistos para o teste de hipóteses é necessário ainda definir os erros envolvidos e as regiões de rejeição e de aceitação.

Para ilustrar estes conceitos será suposto o seguinte teste a ser feito: Dispõem-se de duas moedas com aparência idêntica, só que uma $\{M_1\}$ é equilibrada, isto é, $P\{\text{"Cara"}\} = P\{\text{"Coroa"}\} = 50\%$, enquanto que a outra $\{M_2\}$ é viciada de tal forma que favorece cara na proporção de 80%, ou seja, $P\{\text{"Cara"}\} = 80\%$ enquanto que $P\{\text{"Coroa"}\} = 20\%$. Supõem-se que uma das moedas é lançada e que com base na variável $X = \text{número de caras}$, deve-se decidir qual delas foi lançada. Neste caso o teste a ser feito envolve as seguintes hipóteses:

H_0 : A moeda lançada é a equilibrada $\{M_1\}$, ou seja, $p = 50\%$

H_1 : A moeda lançada é a viciada $\{M_2\}$, ou seja $p = 80\%$, onde “ p ” é a proporção de caras.

Tem-se que tomar a decisão de apontar qual foi a moeda lançada, baseado apenas em uma amostra, por exemplo 5 lançamentos, de uma população infinita de lançamentos possíveis. A decisão, é claro, estará sujeita a erros, pois se está tomando a decisão em condições de incerteza.

A decisão será baseada nas distribuições amostrais das duas moedas. A tabela 01 mostra as probabilidades de se obter os valores: 0, 1, 2, 3, 4 e 5, da variável $X = \text{número de caras}$, em 5 lançamentos de cada uma das moedas.

Tabela 01 - Probabilidades de se obter cara em 5 lançamentos de uma moeda

| x | $P\{X = x\} \text{ sob } H_0$ | $P\{X = x\} \text{ sob } H_1$ |
|--------------|--|--|
| 0 | $1/32 \rightarrow 3,125\%$ | $1/3125 \rightarrow 0,032\%$ |
| 1 | $5/32 \rightarrow 15,625\%$ | $20/3125 \rightarrow 0,640\%$ |
| 2 | $10/32 \rightarrow 31,250\%$ | $160/3125 \rightarrow 5,120\%$ |
| 3 | $10/32 \rightarrow 31,250\%$ | $640/3125 \rightarrow 20,480\%$ |
| 4 | $5/32 \rightarrow 15,625\%$ | $1280/3125 \rightarrow 40,960\%$ |
| 5 | $1/32 \rightarrow 3,125\%$ | $1024/3125 \rightarrow 32,768\%$ |
| Total | 1 \rightarrow 100% | 1 \rightarrow 100% |

falsa. A decisão de que a hipótese é provavelmente verdadeira ou falsa é tomada com base em distribuições de probabilidade denominadas de “distribuições amostrais”. Em estatística trabalha-se com dois tipos de hipótese.

A **hipótese nula** é a hipótese de igualdade. Esta hipótese é denominada de hipótese de nulidade e é representada por H_0 (lê-se h zero). A hipótese nula é normalmente formulada com o objetivo de ser rejeitada. A rejeição da hipótese nula envolve a aceitação de outra hipótese denominada de **alternativa**. Esta hipótese é a definição operacional da hipótese de pesquisa que se deseja comprovar. A natureza do estudo vai definir como deve ser formulada a hipótese alternativa. Por exemplo, se o teste é do tipo paramétrico, onde o parâmetro a ser testado é representado por θ , então a hipótese nula seria: $H_0 : \theta = \theta_0$ e as hipóteses alternativas seriam:

$H_1 : \theta = \theta_1$ (Hipótese alternativa simples) ou

$H_1 : \theta \neq \theta_0 ; \theta > \theta_0$ ou $\theta < \theta_0$. (Hipóteses alternativas compostas)

No primeiro caso, $H_1 : \theta \neq \theta_0$, diz-se que o teste é bilateral (ou bicaudal), se $H_1 : \theta > \theta_0$, diz-se que o teste é unilateral (ou unicaudal) à direita e se $H_1 : \theta < \theta_0$, então, diz-se que o teste é unilateral (ou unicaudal) à esquerda.

A escolha do teste estatístico

Existem inúmeros testes estatísticos tanto paramétricos quanto não paramétricos.

Alguns itens devem ser levados em conta na escolha da prova estatística para determinada situação. A maneira como a amostra foi obtida, a natureza da população da qual se extraiu a amostra e o tipo de mensuração ou escala empregado nas definições operacionais das variáveis envolvidas, isto é, o conjunto de valores numéricos e ainda o tamanho da amostra disponível.

Uma vez determinados a natureza da população e o método de amostragem ficará estabelecido o modelo estatístico. Associado a cada teste estatístico tem-se um modelo estatístico e condições de mensuração, o teste é válido sob as condições especificadas no modelo e pelo nível da escala de mensuração. Nem sempre é possível verificar se todas as condições do modelo foram satisfeitas e neste caso tem-se que admitir que estas condições foram satisfeitas. Estas condições do modelo estatístico são denominadas *suposições* ou *hipóteses* do teste. Qualquer decisão tomada através de um teste estatístico somente terá validade se as condições do modelo forem válidas.

É óbvio que quanto mais fracas forem as suposições do modelo mais gerais serão as conclusões. No entanto, as provas mais poderosas, isto é, as que apresentam maior

Metodologia do teste de hipóteses.

Nas ciências do comportamento, efetua-se levantamentos a fim de determinar o grau de aceitação de hipóteses baseadas em teorias do comportamento. Formulada uma determinada hipótese particular é necessário coletar dados empíricos e com base nestes dados decide-se então sobre a validade ou não da hipótese. A decisão sobre a hipótese pode levar a rejeição, revisão ou aceitação da teoria que a originou.

Para se chegar a conclusão que uma determinada hipótese deverá ser aceita ou rejeitada, baseado em um particular conjunto de dados, é necessário dispor de um processo objetivo que permita decidir sobre a veracidade ou falsidade de tal hipótese.

A objetividade deste processo deve ser baseada na informação proporcionada pelos dados, e como estes dados, em geral, envolvem apenas parte da população que se pretende atingir, no risco que se está disposto a correr de que a decisão tomada não esteja correta.

A metodologia para a decisão sobre a veracidade ou falsidade de uma determinada hipótese envolve algumas etapas.

1. Definir a hipótese de igualdade (H_0).
2. Escolher a prova estatística (com o modelo estatístico associado) para tentar rejeitar H_0 .
3. Definir o nível de significância (α) e um tamanho de amostra (n).
4. Determinar (ou supor determinada) a distribuição amostral da prova estatística sob a hipótese de nulidade.
5. Definir a região de rejeição.
6. Calcular o valor da prova estatística, utilizando os valores obtidos na(s) amostra(s).
Se tal valor estiver na região de rejeição, rejeitar, então a hipótese nula, senão a decisão será que a hipótese nula não poderá ser rejeitada ao nível de significância determinado.

As hipóteses

Uma hipótese estatística é uma suposição ou afirmação que pode ou não ser verdadeira, relativa a uma ou mais populações. A veracidade ou falsidade de uma hipótese estatística *nunca* é conhecida com certeza, a menos que, se examine toda a população, o que é impraticável na maior parte das situações.

Desta forma, toma-se uma amostra aleatória da população de interesse e com base nesta amostra é estabelecido se a hipótese é provavelmente verdadeira ou provavelmente

Trabalhando esta desigualdade, segue que:

$$P(P - z_{\alpha/2}\sigma_P < \mu_P < P + z_{\alpha/2}\sigma_P) = P(P - z_{\alpha/2}\sigma_P < \pi < P + z_{\alpha/2}\sigma_P) = 1 - \alpha$$

Que é o intervalo procurado. Assim o intervalo de confiança (probabilidade) de “1 - α ” para a proporção “P” de uma população é dado por:

$$\left[P - z_{\alpha/2} \sqrt{\frac{\pi(1-\pi)}{n}}; P + z_{\alpha/2} \sqrt{\frac{\pi(1-\pi)}{n}} \right]$$

Observando-se a expressão acima pode-se perceber que o intervalo de confiança para a proporção populacional π , depende dele mesmo, isto é, é necessário calcular o erro amostral que está expresso em função de π . Como o objetivo é estimar este valor, evidentemente ele não é conhecido. Assim é necessário utilizar, sua estimativa $\hat{\sigma}_P$, isto

é, é necessário substituir π por P na expressão $\sigma_P = \sqrt{\frac{\pi(1-\pi)}{n}}$. Desta forma o intervalo

acima ficará:

$$\left[P - z_{\alpha/2} \sqrt{\frac{P(1-P)}{n}}; P + z_{\alpha/2} \sqrt{\frac{P(1-P)}{n}} \right]$$

Onde:

P é a estimativa por ponto da proporção populacional π .

$\hat{\sigma}_P = \sqrt{\frac{P(1-P)}{n}}$ é uma estimativa do erro padrão, isto é, do desvio padrão amostral.

$z_{\alpha/2}$ é o valor da distribuição normal padrão cuja área à direita é igual a $\alpha/2$. É o valor de Z tal que: $P(Z > z_{\alpha/2}) = \alpha/2$, ou então: $\Phi(-z_{\alpha/2}) = \alpha/2$.

TESTES DE HIPÓTESES

Generalidades

Um dos principais assuntos da Estatística moderna é a *inferência estatística*. A inferência estatística é dividida em dois grandes tópicos: a estimação de parâmetros de uma população e os testes de hipóteses.

No desenvolvimento dos métodos da estatística moderna, as primeiras técnicas de inferência que apareceram foram as que faziam diversas hipóteses sobre a natureza da população da qual se extraíram os dados. Como os valores relacionados com a população são denominados “parâmetros”, tais técnicas estatísticas foram denominadas de paramétricas.

(b) Desvio padrão populacional (σ) desconhecido

Quando o desvio padrão da população (σ) é desconhecido é necessário utilizar sua estimativa “S”. Só que ao substituir-se o desvio padrão populacional pelo sua estimativa no quociente:

$$\frac{\bar{X} - \mu}{S/\sqrt{n}}$$

Não se terá mais uma normal padrão. De fato, conforme demonstrado pelo estatístico inglês W. S. Gosset, conhecido por “Student” o comportamento do quociente segue uma distribuição simétrica em torno de zero, porém com uma variabilidade maior do que a da normal padrão. A distribuição do quociente acima é conhecida como distribuição “t” de Student.

Neste caso, o intervalo de confiança com probabilidade “1 - α ” para a média será:

$$\left[\bar{X} - t_{\alpha/2} S/\sqrt{n}; \bar{X} + t_{\alpha/2} S/\sqrt{n} \right]$$

Onde:

\bar{X} é a estimativa por ponto da média da população;

S é o desvio padrão da amostra, e uma estimativa do desvio padrão da população σ e,

$t_{\alpha/2}$ é o valor da distribuição t cuja área à direita é igual a $\alpha/2$, isto é, é o valor de t tal que:

$$P(t > t_{\alpha/2}) = \alpha/2, \text{ ou então: } P(-t_{\alpha/2} < t < t_{\alpha/2}) = 1 - \alpha.$$

Proporção populacional

Seja P = proporção amostral. Sabe-se que para $n > 30$ a distribuição amostral de P é aproximadamente normal com média $\mu_P = \pi$ e desvio padrão (erro padrão)

$$\sigma_P = \sqrt{\frac{\pi(1-\pi)}{n}}. \text{ Pode-se então utilizar a curva normal para estabelecer os limites para o}$$

intervalo de confiança.

Lembrando que o que se quer é um intervalo que contenha o parâmetro populacional π com probabilidade “1 - α ” então tem-se:

$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$, onde $z_{\alpha/2}$ é o valor da normal padrão com área à direita é igual a $\alpha/2$.

$$\text{Mas } Z = (P - \mu_P) / \sigma_P$$

Então substituindo na expressão acima vem:

$$P(-z_{\alpha/2} < (P - \mu_P) / \sigma_P < z_{\alpha/2}) = 1 - \alpha.$$

Média da população

(a) Desvio padrão populacional (σ) conhecido

O intervalo de confiança para a média (μ) de uma população é construído em torno da estimativa pontual \bar{X} . Para construir este intervalo fixa-se uma probabilidade “ $1 - \alpha$ ” de que o intervalo construído contenha o parâmetro populacional. Desta forma, “ α ” será a probabilidade de que o intervalo obtido não contenha o valor do parâmetro, isto é, “ α ” será a probabilidade de erro. Sabe-se que a média da amostra tem distribuição normal de média μ e desvio padrão σ/\sqrt{n} se a população de onde for extraída a amostra for normal (ou se a amostra for superior a 30 e retirada de qualquer população) de média μ e de desvio padrão σ , pode-se então utilizar a curva normal para estabelecer os limites para o intervalo de confiança.

Lembrando que o que se quer é um intervalo que contenha o parâmetro populacional μ com probabilidade “ $1 - \alpha$ ” tem-se então:

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$$

Onde $z_{\alpha/2}$ é o valor da normal padrão com área à direita é igual a $\alpha/2$. Mas $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$

substituindo na expressão acima vem:

$$P\left(-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1 - \alpha$$

Trabalhando esta desigualdade, segue que:

$$P(\bar{X} - z_{\alpha/2} \sigma/\sqrt{n} < \mu < \bar{X} + z_{\alpha/2} \sigma/\sqrt{n}) = 1 - \alpha$$

Que é o intervalo procurado. Assim o intervalo de confiança (probabilidade) de “ $1 - \alpha$ ” para a média de uma população é dado por:

$$\left[\bar{X} - z_{\alpha/2} \sigma/\sqrt{n}; \bar{X} + z_{\alpha/2} \sigma/\sqrt{n}\right] \text{ onde:}$$

\bar{X} é a estimativa por ponto da média da população.

σ é o desvio padrão da população e

$z_{\alpha/2}$ é o valor da distribuição normal padrão cuja área à direita é igual a $\alpha/2$, isto é, é o valor de Z tal que: $P(Z > z_{\alpha/2}) = \alpha/2$, ou então: $\Phi(-z_{\alpha/2}) = \alpha/2$ ²

² $\Phi(z) = F(z)$, onde F se refere a distribuição Normal padrão.

$300/400 = 75\%$ é uma estimativa por ponto do percentual de pessoas da cidade que acham a administração boa ou ótima. Esta mesma estimativa poderia ser enunciado como de: 70% a 80% das pessoas da cidade acham a administração boa ou ótima. Neste caso, teríamos uma estimativa por intervalo da proporção. Note-se que o centro do intervalo é o valor “75%” da estimativa pontual.

Propriedades dos estimadores

Seja X uma população com um parâmetro de interesse θ e seja (X_1, X_2, \dots, X_n) uma amostra aleatória simples extraída desta população. Seja $\hat{\theta}$ um estimador do parâmetro θ . Então:

- (i) Se $E(\hat{\theta}) = \theta$ se dirá que $\hat{\theta}$ é um estimador não-tendencioso ou não viciado do parâmetro populacional θ . Neste caso, a média do estimador $\hat{\theta}$ é o parâmetro populacional θ , ou ainda, pode-se dizer que o estimador varia em torno do parâmetro populacional.
- (ii) Se $\hat{\theta}$ é um estimador não tendencioso de um parâmetro θ , se dirá que $\hat{\theta}$ é consistente se à medida que o tamanho da amostra aumenta a variabilidade do estimador diminui, isto é, as observações vão ficando cada vez mais concentradas em torno do parâmetro na medida em que a amostra vai ficando cada vez maior. Em símbolos:

$$\lim_{n \rightarrow \infty} V(\hat{\theta}) = 0$$

Estimação por ponto

Seja X uma população com média μ , desvio padrão σ e com uma proporção π e seja (X_1, X_2, \dots, X_n) uma amostra aleatória simples extraída desta população. Então:

- (a) \bar{X} é um estimador não-tendencioso e consistente da média da população μ .
- (b) P é um estimador não-tendencioso e consistente da proporção populacional π .
- (c) S^2 é estimador não-tendencioso e consistente da variância da população σ^2 , a menos que a extração seja **sem** reposição de população finita. Neste caso, o estimador é $\hat{S}^2 = \frac{N-1}{N} S^2$.

Estimação por intervalo

O estimador por ponto não permite ter uma idéia do erro cometido ao se fazer a estimativa do parâmetro. Para que se possa associar uma confiança (probabilidade) a uma estimativa é necessário construir um intervalo em torno da estimativa por ponto. Este intervalo é construído baseado na distribuição amostral do estimador.

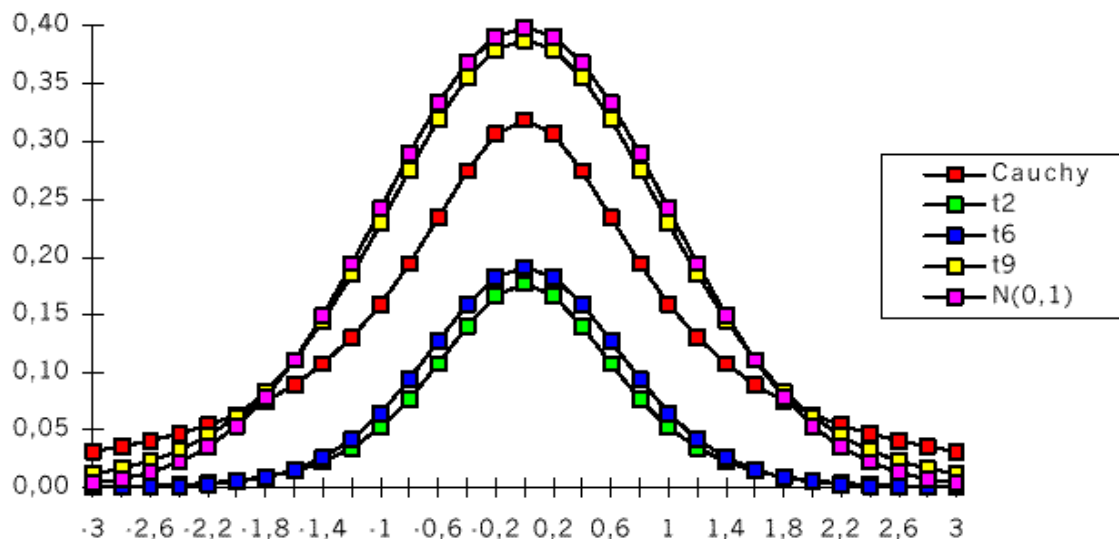


Figura 10- Densidades t com 1, 2, 6 e 9 graus de liberdade e densidade $N(0, 1)$

Como mostram os gráficos, a densidade t é simétrica em relação a $t = 0$. A distribuição t é completamente determinada pelo parâmetro k , o número de graus de liberdade (que é o mesmo número de graus de liberdade que aparece na densidade Quiquadrado usada para gerar a densidade t). Também, à medida que aumentam os graus de liberdade, a densidade t se aproxima de uma $N(0,1)$.

Teorema

Sejam X_1, X_2, \dots, X_n uma amostra aleatória da densidade $N(\mu, \sigma^2)$. Sejam \bar{X} e S^2 respectivamente a média e variância amostral. Então a estatística:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{\sqrt{n}(\bar{X} - \mu)}{S}$$

Tem distribuição t com $n - 1$ graus de liberdade.

ESTIMAÇÃO

A inferência estatística tem por objetivo fazer generalizações sobre uma população com base em valores amostrais. A inferência pode ser feita estimando os parâmetros:

- (a) Por ponto e
- (b) Por intervalo.

A estimação por ponto é feita através de um único valor, enquanto que a estimação por intervalo fornece um intervalo de valores em torno do valor da estimativa pontual.

Exemplo:

Uma amostra aleatória simples de 400 pessoas de uma cidade é extraída e 300 respondem que acham a administração municipal boa ou ótima. Então o valor $p =$

Teorema

Sejam X_1, X_2, \dots, X_n uma amostra aleatória da distribuição $N(\mu, \sigma^2)$. Seja σ^2 a variância amostral. Então:

$$E(S^2) = \sigma^2$$

$$V(S^2) = \frac{2\sigma^4}{n-1}$$

As distribuições t e F

As distribuições t e F são importantes nos contextos de estimação, intervalos de confiança e testes de hipóteses para amostras Normais. A seguir apresentamos estas duas distribuições. O fundamental aqui não é saber as fórmulas das densidades de cor, e sim identificar como estas densidades surgem a partir de outras densidades, como a Normal e a quiquadrado.

Teorema (a "gênese" da distribuição t)

Sejam $Z \sim N(0,1)$ e $V \sim \chi_k^2$ variáveis aleatórias independentes. Defina uma nova variável aleatória T como:

$$T = \frac{Z}{\sqrt{V/k}}$$

Então T tem distribuição t de Student com k graus de liberdade, e sua densidade é dada por:

$$f(t) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi}\Gamma\left(\frac{k}{2}\right)\left(1 + \frac{t^2}{k}\right)^{\frac{k+1}{2}}}$$

Onde t é um número real qualquer.

$$2) \text{VAR}(\bar{X}) = \frac{\sigma^2}{n}$$

3) Se n é grande, pelo teorema central do limite podemos concluir que:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

é aproximadamente $N(0,1)$.

Note que, neste caso, nada é dito a respeito da distribuição de X . Apenas a sua média e variância são conhecidas, e são funções da média e variância de cada X_i . A princípio a distribuição de X poderia ser uma coisa estranha, que não tem nada a ver com a distribuição original de cada X_i . No entanto, se o tamanho da amostra é grande podemos concluir que a distribuição de X , devidamente escalonada, é aproximadamente $N(0,1)$. O próximo teorema refere-se à distribuição do máximo e do mínimo de uma amostra.

Teorema

Sejam X_1, X_2, \dots, X_n uma amostra aleatória de uma distribuição contínua qualquer com densidade $f()$ e função de distribuição $F()$. Sejam $X_{(1)}$ e $X_{(n)}$ respectivamente, o mínimo e o máximo da amostra. Então as densidades de $X_{(1)}$ e $X_{(n)}$ são dadas por:

1) Densidade do mínimo

$$g_1(x) = nf(x)(1 - F(x))^{n-1}$$

2) Densidade do máximo

$$g_1(x) = nf(x)(F(x))^{n-1}$$

Teorema

Sejam X_1, X_2, \dots, X_n uma amostra aleatória da distribuição $N(\mu, \sigma^2)$. Seja S^2 a variância amostral, dada por:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

Então:

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2}$$

Tem distribuição Quiquadrado com $(n-1)$ graus de liberdade.

A partir deste teorema podemos deduzir facilmente a média e variância de σ^2 .

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

3) Desvio padrão amostral

$$S = \sqrt{S^2} = \sqrt{\frac{1}{n-1} \sum (X_i - \bar{X})^2}$$

4) Mínimo da amostra

$$X_{(1)} = \min(X_1, X_2, \dots, X_n)$$

5) Máximo da amostra

$$X_{(n)} = \max(X_1, X_2, \dots, X_n)$$

6) Amplitude da amostra

$$A = X_{(n)} - X_{(1)}$$

7) *k*ésima estatística de ordem

É o *k*ésimo elemento da amostra ordenada. Por exemplo, $X_{(2)}$ é o segundo menor elemento da amostra X_1, X_2, \dots, X_n .

Um dos nossos objetivos aqui é desenvolver as distribuições de estatísticas obtidas a partir de uma amostra aleatória da distribuição Normal.

O próximo teorema refere-se à média amostral de uma amostra aleatória da densidade Normal.

Teorema

Sejam X_1, X_2, \dots, X_n uma amostra aleatória da distribuição $N(\mu, \sigma^2)$. Seja \bar{X} a média amostral. Então:

$$\bar{X} = N\left(\mu, \frac{\sigma^2}{n}\right)$$

A demonstração do teorema é trivial, e segue das propriedades da função geradora de momentos.

Este teorema pode ser generalizado para uma amostra aleatória de uma distribuição qualquer.

Teorema

Sejam X_1, X_2, \dots, X_n uma amostra aleatória de uma distribuição qualquer tal que $E(X_i) = \mu$ e $\text{VAR}(X_i) = \sigma^2$. Seja \bar{X} a média amostral. Então:

$$1) E(\bar{X}) = \mu$$

Para ganhar informação sobre os parâmetros desconhecidos de uma distribuição de probabilidade usamos um conjunto de variáveis aleatórias independentes e identicamente distribuídas. Isto equivale a repetir a experiência aleatória que está sendo descrita pelo modelo em questão n vezes, em condições idênticas e de maneira independente. A partir dos valores observados das variáveis X_1, X_2, \dots, X_n calcularemos funções que nos permitirão aprender sobre os parâmetros desconhecidos do modelo. Estas funções serão chamadas de "estatísticas".

Definição (estatística)

Seja X_1, X_2, \dots, X_n uma a.a. de uma variável aleatória X . Sejam x_1, x_2, \dots, x_n os valores observados de X_1, X_2, \dots, X_n . Seja $Y = h(X_1, X_2, \dots, X_n)$ uma função apenas das variáveis X_1, X_2, \dots, X_n . Y é chamado de "estatística".

Note que uma estatística não é função de parâmetros desconhecidos, ela só envolve as variáveis na amostra aleatória.

Por definição, qualquer estatística Y é uma variável aleatória, e tem uma distribuição de probabilidade que depende da distribuição de X_1, X_2, \dots, X_n .

O nosso problema então é encontrar estatísticas que sirvam como bons estimadores pontuais de parâmetros desconhecidos. Também é importante definir critérios que nos permitam dizer que uma estatística é "melhor" que outra para estimar um dado parâmetro.

De uma maneira geral, as estatísticas devem conter "toda" a informação presente numa amostra. Se não fosse assim, não valeria a pena calcular uma estatística, a gente simplesmente usaria uma única observação da amostra. Este acréscimo de informação representado pela uso de uma estatística (ao invés de uma única observação) geralmente se traduz por uma considerável redução na variância. Por exemplo, a variância da média amostral é igual à variância de cada observação dividida pelo tamanho da amostra. Quanto maior o tamanho da amostra, menor é a variância da média amostral, isto é, mais "precisa" é a média amostral.

As estatísticas mais famosas

Sejam X_1, X_2, \dots, X_n uma amostra aleatória de uma distribuição qualquer. As estatísticas mais comuns, calculadas a partir desta amostra são:

1) Média amostral

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

2) Variância amostral

etc... eram adequados. Todos estes modelos referem-se a distribuições de probabilidade que envolvem parâmetros, que até agora foram supostos conhecidos. Para que as probabilidades associadas a eventos sejam calculadas é necessário conhecer o valor destes parâmetros.

No estudo das probabilidades realizado até agora, o nosso objetivo era calcular a probabilidade de eventos pré-especificados. De agora em diante teremos um novo objetivo. A partir de uma amostra de uma distribuição de probabilidade especificada pretendemos aprender alguma coisa sobre os parâmetros da distribuição, isto é, estaremos interessados em estimar os parâmetros da distribuição de probabilidade. Esta é a grande diferença entre Probabilidade e Estatística. No estudo de Probabilidade estamos interessados em definir modelos que possam ser aplicados a situações reais. Estes modelos envolvem distribuições de probabilidade totalmente conhecidas, isto é, não apenas a forma da densidade, mas também os seus parâmetros são conhecidos. No estudo da Estatística supõe-se que o modelo probabilístico é conhecido, isto é, sabe-se qual a distribuição de probabilidade que modela a situação real, mas os parâmetros desta distribuição são desconhecidos, e devem ser estimados a partir dos dados. O nosso objetivo em Estatística é descobrir alguma coisa sobre os parâmetros desconhecidos de uma distribuição de probabilidade. Os mecanismos mais usuais para "inferir" alguma coisa sobre estes parâmetros são:

1. Estimação pontual -o objetivo é "chutar" os valores do parâmetro desconhecido.
2. Estimação por intervalos -o objetivo é encontrar um intervalo que contenha o parâmetro de interesse com uma probabilidade especificada.
3. Testes de hipóteses -o objetivo é criar conjecturas sobre os valores possíveis do parâmetro e verificar se estas conjecturas são muito ou pouco prováveis (isto é, testar as hipóteses).

Todos estes procedimentos são baseados na noção de **amostra aleatória**.

Definição (amostra, ou amostra aleatória)

Uma amostra aleatória é um conjunto de variáveis aleatórias **independentes e identicamente distribuídas** (iid).

Notação: a.a. = amostra aleatória

O que se faz na prática?

Obviamente a definição não se aplica se algum dos $E(X^k)$ é infinito.

Definição: k ésimo momento central ou k ésimo momento em torno da média)

O k ésimo momento central da variável aleatória X é definido como:

$$E(X^k) = \begin{cases} \sum_x (x - \mu)^k f(x), & \text{caso discreto} \\ \int_{-\infty}^{+\infty} (x - \mu)^k f(x) dx, & \text{caso contínuo} \end{cases}$$

Onde $k = 1, 2, 3, \dots$

Logo, a média e a variância são apenas casos particulares de momentos. A média é o primeiro momento (isto é, $\mu = E(X)$) e a variância é o segundo momento central, ou seja: $E(X - \mu)^2$.

A notação $E(\dots)$ indica um valor esperado, e pode ser estendida para funções mais gerais que X^k ou $(X - \mu)^k$.

Definição (valor esperado de uma função de uma variável aleatória)

Seja X uma variável aleatória com densidade $f(x)$ e seja $u(X)$ uma função qualquer tal que:

$$E[u(X)] = \begin{cases} \sum_x u(X) f(x), & \text{caso discreto} \\ \int_{-\infty}^{+\infty} u(X) f(x) dx, & \text{caso contínuo} \end{cases}$$

Formula alternativa para o cálculo da Variância

$$\begin{aligned} V(X) &= E[X - E(X)]^2 \\ &= E\{X^2 - 2XE(X) + [E(X)]^2\} \\ &= E(X^2) - E[2XE(X)] + E\{[E(X)]^2\} \\ &= E(X^2) - 2E(X)E(X) + [E(X)]^2 \\ &= E(X^2) - [E(X)]^2 \end{aligned}$$

Esta fórmula é válida para qualquer variável aleatória X (contínua ou discreta), desde que a média de X seja finita.

AMOSTRAS ALEATÓRIAS E DISTRIBUIÇÕES AMOSTRAIS

A partir de agora mudamos um pouco o enfoque do curso. Até agora o que fizemos foi desenvolver modelos probabilísticos que se adequavam a situações reais. Por exemplo, indicamos quando os modelos Binomial, Poisson, Exponencial, Normal, Uniforme,

A média de uma variável aleatória representa uma medida de tendência central da distribuição de probabilidade desta variável aleatória.

A variância de uma variável aleatória é uma medida da dispersão da distribuição de probabilidade, definida como:

$$\sigma^2 = V(X) = \begin{cases} \sum_x (x - \mu)^2 f(x), & \text{caso discreto} \\ \int_{-\infty}^{+\infty} (x - \mu)^2 f(x) dx, & \text{caso contínuo} \end{cases}$$

Onde novamente $f(x)$ representa a densidade de probabilidade (discreta ou contínua) da variável aleatória X e μ é a média da variável aleatória. A variância é o segundo momento em torno da média, e corresponde ao momento de inércia em Mecânica. Da própria definição segue que a variância é uma quantidade sempre maior ou igual a zero.

Definição (desvio padrão)

O desvio padrão de uma variável aleatória é a raiz quadrada positiva da sua variância, e denotado por σ .

O desvio padrão é expresso nas mesmas unidades que a variável aleatória, e a variância é dada nas unidades da variável aleatória ao quadrado. Logo, se a variável aleatória é medida em metros, o desvio padrão também está em metros e a variância em metros quadrados. Um valor pequeno do desvio padrão indica que existe pouca dispersão em torno da média. Se o desvio padrão é grande, os valores da variável aleatória estão muito dispersos em torno da média.

A média e a variância são casos particulares do que chamamos de "momentos" de uma distribuição de probabilidade. Os momentos de uma distribuição servem para caracterizar esta distribuição, não apenas no que se refere à sua centralidade e dispersão, mas também com relação a outras características, como a simetria ou assimetria da densidade de probabilidade.

Definição: k-ésimo momento

O k -ésimo momento da variável aleatória X é definido como:

$$E(X^k) = \begin{cases} \sum_x x^k f(x), & \text{caso discreto} \\ \int_{-\infty}^{+\infty} x^k f(x) dx, & \text{caso contínuo} \end{cases}$$

Onde $k = 1, 2, 3, \dots$

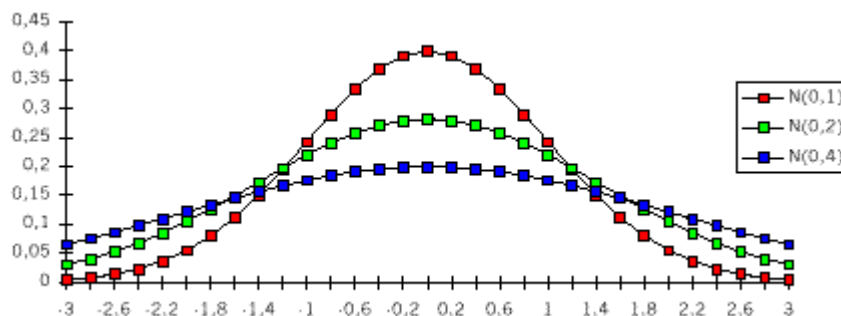


Figura 9- Distribuições Normais com média zero e variâncias 1, 2 e 4

Note que o máximo das densidades é encontrado quando $x = 0$, isto é, quando x é igual à média da distribuição. Isto vale para qualquer distribuição Normal: o máximo de $f(x)$ é obtido fazendo-se $x = \mu$, onde μ é a média da Normal. Também, quanto maior o valor da variância σ^2 , mais "espalhada" é a distribuição.

• Propriedades da Distribuição Normal

- 1) $f(x)$ dada pela expressão acima integra a 1.
- 2) $f(x) \geq 0$ sempre.
- 3) Os limites de $f(x)$ quando x tende a $+\infty$ e $-\infty$ são iguais a zero.
- 4) A densidade $N(\square, \square^2)$ é **simétrica em torno de** \square , ou seja:

$$f(\square + x) = f(\square - x)$$

- 5) O valor máximo de $f(x)$ ocorre em $x = \square$
- 6) Os pontos de inflexão de $f(x)$ são $x = \square + \square$ e $x = \square - \square$

Momentos de uma distribuição de probabilidade

A seguir definimos alguns dos momentos de distribuições de probabilidade contínuas e discretas. Momentos são quantidades que nos dão uma idéia da tendência central, dispersão e assimetria de uma densidade de probabilidades.

Definição (média e variância)

A média (ou valor esperado, primeiro momento de uma variável aleatória) é definida como:

$$\mu = E(X) = \begin{cases} \sum_x x f(x), & \text{caso discreto} \\ \int_{-\infty}^{+\infty} x f(x) dx, & \text{caso contínuo} \end{cases}$$

Onde o somatório refere-se a todos os valores de X quando X é uma variável discreta. Quando X é uma variável contínua a média é calculada pela integral acima, onde $f(x)$ representa a densidade de probabilidade da variável X .

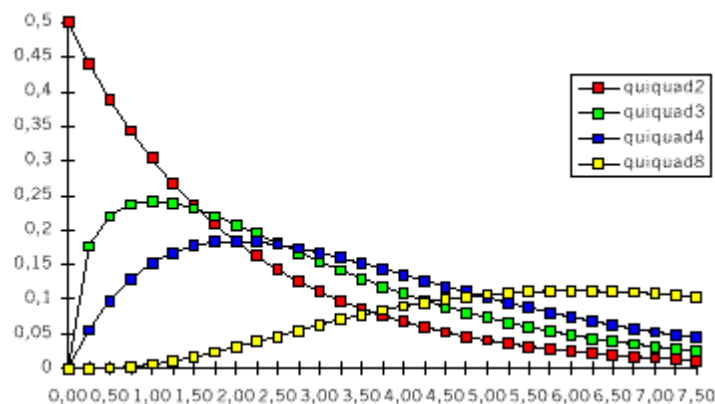


Figura 8- Distribuições Qui-quadrado com 2, 3, 4 e 8 graus de liberdade.

Distribuição Normal (Gaussiana)

A distribuição Normal é talvez a mais importante das distribuições de probabilidade, por razões que ficarão claras ao longo deste curso. Erros de mensuração de fenômenos físicos ou econômicos são freqüentemente modelados pela distribuição Normal, mas esta não é a única aplicação desta densidade. Por exemplo, a distribuição dos pesos, alturas e QI's das pessoas numa população também já foram modelados com sucesso por esta distribuição. A distribuição Normal tem a forma de um sino, e possui dois parâmetros, μ e σ^2 .

A distribuição Normal é também chamada de Gaussiana em homenagem ao matemático Carl Friederich Gauss (1777 - 1855), que a utilizou pela primeira vez na modelagem de erros de medida. A distribuição Normal também funciona como uma boa aproximação para outras densidades. Por exemplo, sob algumas condições pode-se provar que a densidade Binomial pode ser aproximada pela Normal.

Definição - Densidade Normal com média μ e variância σ^2

Seja X uma variável aleatória contínua definida nos números reais. Dizemos que X tem densidade Normal com média μ e variância σ^2 se a densidade de X é:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Notação: $X \sim N(\mu, \sigma^2)$

Note que o segundo parâmetro (σ^2) nesta notação é a variância de X . A seguir exibimos gráfico das distribuições Normais com média zero e variâncias 1, 2 e 4.

Densidade Gama

Seja X uma variável aleatória contínua definida no intervalo $(0, \infty)$. Dizemos que X tem densidade Gama com parâmetros α e β , e escrevemos $X \sim \text{Gama}(\alpha, \beta)$ se a densidade de X é:

$$f(x) = \begin{cases} \frac{\beta^\alpha x^{\alpha-1}}{\Gamma(\alpha)} e^{-\beta x}, & \text{onde } x \geq 0 \\ 0 & \text{se } x < 0 \end{cases}$$

Os parâmetros α e β são números reais positivos. α é conhecido como parâmetro de forma, e β é o parâmetro de escala.

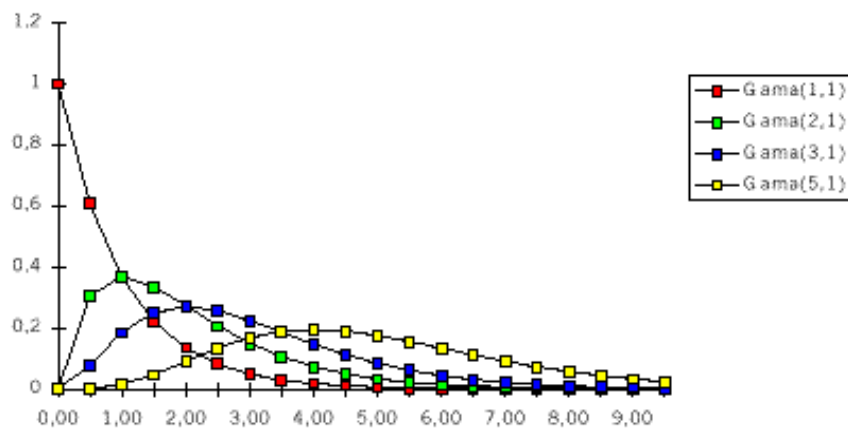


Figura 7- Densidades Gama: Gama(1,1), Gama(1,2), Gama(1,3) e Gama(1,5)

Densidade Quiquadrado com k graus de liberdade)

Seja X uma variável aleatória contínua e positiva com densidade dada por:

$$f(x) = \frac{1}{2^{\frac{n}{2}} \Gamma\left(\frac{k}{2}\right)} x^{\frac{n-2}{2}} e^{-\frac{x}{2}} \quad \text{onde } x > 0$$

Tem densidade Quiquadrado com $(n - 1)$ graus de liberdade, e escrevemos:

$$X \sim \chi_k^2$$

A densidade Quiquadrado com k graus de liberdade é apenas um caso particular da densidade Gama. Na verdade:

$$\chi_k^2 = \text{Gama}\left(\alpha = \frac{k}{2}, \beta = \frac{1}{2}\right)$$

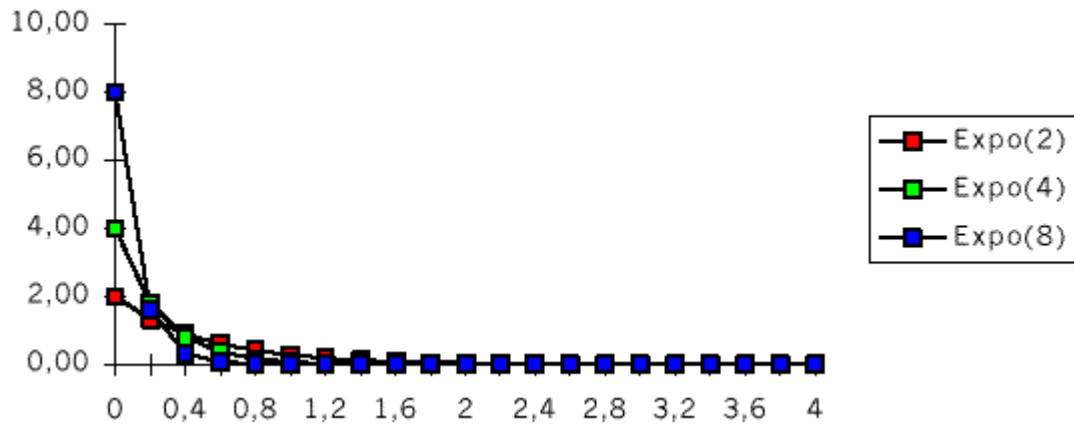


Figura 5- Densidades exponenciais

O próximo gráfico exibe a função de distribuição de uma variável aleatória com parâmetros $\lambda = 1$ e $\lambda = 2$.

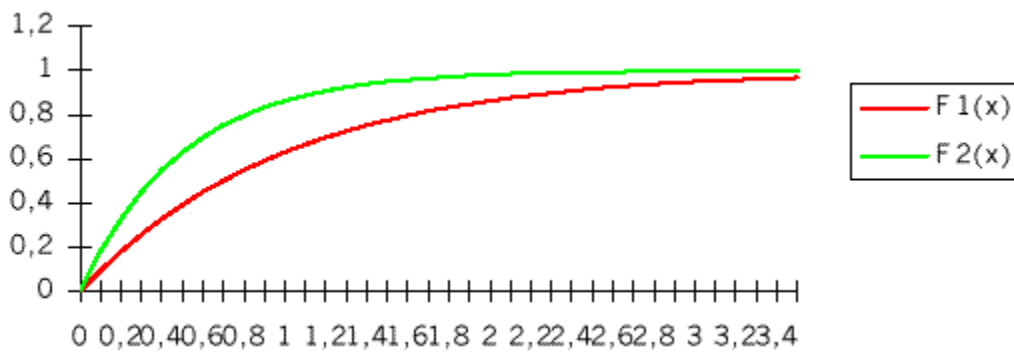


Figura 6- Funções de Distribuição -densidades Expo(1) e Expo(2)

Definição: Função Gama

Seja a um número real maior que zero, não necessariamente inteiro. A função Gama com argumento α é definida por:

$$\Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1} e^{-t} dt$$

Propriedades da Função Gama

$$1) \Gamma(n) = (n-1)\Gamma(n-1) \text{ para } n > 1$$

A demonstração deste fato usa integração por partes.

$$2) \Gamma(n) = (n-1)! \text{ se } n \text{ é inteiro } > 1$$

$$3) \Gamma(1) = 0! = 1$$

$$4) \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$$

Uma variável aleatória X tem densidade Uniforme no intervalo (a,b) , e escrevemos $X \sim \text{Unif}(a,b)$ se a sua densidade é:

$$f(x) = \begin{cases} \frac{1}{b-a}, & x \in (a,b) \\ 0, & x \notin (a,b) \end{cases}$$

Uma variável aleatória X com densidade Uniforme no intervalo (a,b) tem a seguinte propriedade: qualquer subintervalo de comprimento d localizado dentro do intervalo (a,b) tem a mesma probabilidade.

A função de distribuição de uma variável aleatória $\text{Unif}(a,b)$ é:

$$F(x) = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & x \in [a,b] \\ 1, & x > b \end{cases}$$

Densidade Exponencial

Uma variável aleatória com densidade Exponencial é usada para modelar tempos de duração de equipamentos. Na verdade, existem densidades mais apropriadas para modelar este fenômeno, pois, como veremos mais tarde, a densidade Exponencial não leva em conta o desgaste do equipamento ao longo do tempo. A densidade Exponencial é definida para variáveis contínuas e maiores que zero, e depende de um parâmetro positivo, λ .

Notação: $X \sim \text{Expo}(\lambda)$

A densidade Exponencial é dada pela fórmula:

$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0$$

E $f(x) = 0$ se $x < 0$. Note que $\lambda > 0$ sempre.

A função de distribuição é dada por:

$$F(x) = \begin{cases} 0, & x < 0 \\ \int_0^x \lambda e^{-\lambda u} du = -e^{-\lambda u} \Big|_0^x = 1 - e^{-\lambda x}, & x \geq 0 \end{cases}$$

Note que o limite da função de distribuição quando x tende a $+\infty$ é um!

O próximo gráfico apresenta as densidades Exponenciais com parâmetros $\lambda = 2, 4$ e 8 .

Note que a densidade decai mais rápido quando λ é grande.

$$p_0'(t) = -kp_0(t) \quad (4)$$

Para $x > 0$ pode-se provar que as premissas resultam no seguinte sistema de equações diferenciais:

$$p_x'(t) = -kp_x(t) + kp_{x-1}(t), \text{ onde } x = 1, 2, 3, \dots \quad (5)$$

A solução do sistema dado por (4) e (5) é:

$$p_x(t) = \frac{(kt)^x e^{-kt}}{x!}, \text{ onde } x = 1, 2, 3, \dots$$

Para qualquer intervalo $[0, t]$, se fixarmos t e fizermos $\lambda = kt$ a equação acima reduz-se a:

$$P(X = x) = f(x) = \frac{\lambda^x e^{-\lambda}}{x!}, \text{ onde } x = 1, 2, 3, \dots$$

A densidade acima é a densidade Poisson com parâmetro $\lambda > 0$ e escrevemos:

$X \sim \text{Poisson}(\lambda)$.

Como usar a densidade Poisson na prática

Selecione um intervalo de tempo fixo. Conte o número de ocorrência de um certo evento de interesse neste intervalo. Este número de ocorrências é uma variável discreta com valores possíveis 0, 1, 2, Se o evento é tal que a probabilidade do número de ocorrências no intervalo ser 0 ou 1 é "grande", então o evento pode ser na prática modelado pela distribuição de Poisson. Uma densidade Poisson modela bem eventos "raros", isto é, que não acontecem com grande frequência para qualquer intervalo de tempo fixo.

Por exemplo, o número de automóveis Corsa que entram num estacionamento no Rio de Janeiro num intervalo de 1 hora certamente não é uma variável Poisson, mas o número de Ferraris que entram no estacionamento no mesmo período de tempo deve ser Poisson!

Algumas distribuições contínuas

Densidade Uniforme

A densidade Uniforme serve para modelar o seguinte fenômeno: "escolhe-se um número aleatoriamente num intervalo dado", por exemplo, o intervalo (0,1). A função "random", presente na maioria das linguagens de computador, nada mais é do que um mecanismo para gerar números com distribuição Uniforme no intervalo (0,1).

A derivação da densidade Poisson pode ser feita de duas formas: a primeira está relacionada com o "processo de Poisson" e a segunda surge como uma aproximação da densidade Binomial.

Processo de Poisson

Considere uma seqüência de eventos que ocorrem ao longo do tempo, como o número de carros vermelhos que param num sinal, o número de chamadas telefônicas que chegam numa estação durante um certo intervalo de tempo.

Seja X_t o número de ocorrência no intervalo de tempo $[0, t]$. Claramente X_t é uma variável aleatória discreta com valores possíveis 0, 1, 2, Para derivar a densidade de X_t partimos das seguintes premissas:

Seja Δt um intervalo de tempo pequeno. Então:

1. A probabilidade de exatamente uma ocorrência em um intervalo de tempo Δt é aproximadamente $k\Delta t$.
2. A probabilidade de exatamente zero ocorrências em um intervalo de tempo Δt é aproximadamente $(1 - k)\Delta t$.
3. A probabilidade de duas ou mais ocorrências em um intervalo de tempo Δt é igual a um certo $o(\Delta t)$, onde $o(\Delta t)/\Delta t$ tende a zero à medida que Δt tende a zero. Em outras palavras, a probabilidade de duas ou mais ocorrências em um intervalo de tempo Δt é um valor muito pequeno, e este valor decresce à zero mais rapidamente que o comprimento do intervalo Δt .

Estas três premissas definem o tipo de processo que pode ser chamado de um processo de Poisson.

O parâmetro k acima é um número real > 0 , chamado de **taxa média de ocorrência**.

Para cada instante $t > 0$ seja:

$$P(X_t = x) = p_x(t), \text{ onde } x = 0, 1, 2, \dots$$

Fixando um instante qualquer t e aplicando a segunda premissa nos dá:

$$p_1(t + \Delta t) \cong [1 - k\Delta t]p_0(t)$$

Subtraindo $p_0(t)$ de ambos os lados e dividindo o resultado por Δt leva a:

$$\frac{p_1(t + \Delta t) - p_0(t)}{\Delta t} \cong -kp_0(t)$$

Tomando-se o limite desta última expressão quando Δt tende a zero encontramos, do lado esquerdo, a derivada de $p_0(t)$. Isto nos dá a equação diferencial:

Nota : a escolha de um tipo de resultado como "sucesso" ou "falha" não implica em qualquer julgamento sobre o resultado ser "bom" ou "ruim", é apenas uma questão de nomenclatura. Na verdade, a escolha do que é um "sucesso" ou "falha" depende da questão de interesse ao analisar o problema -o que é "sucesso" numa situação pode ser a "falha" num problema semelhante.

A equação da distribuição binomial é a seguinte,

$$P(X = x) = f(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

Distribuição hipergeométrica

Suponha agora que a amostragem é feita sem reposição e que o tamanho da população não é muito maior que o da amostra, de tal forma que a densidade Binomial não pode ser usada. Sejam N e n respectivamente os tamanhos da população e da amostra, e suponha que na população existem r objetos do tipo A ("sucessos") e $N - r$ objetos do tipo B ("falhas").

Seja X o número de objetos do tipo A na amostra. Então as probabilidades dos diversos valores de X são dadas pela seguinte fórmula:

$$P(X = x) = f(x) = \frac{\binom{r}{x} \cdot \binom{N-r}{n-x}}{\binom{N}{n}}$$

onde $x = 0, 1, 2, \dots, \min(n, r)$.

Esta é a densidade Hipergeométrica, usada para calcular probabilidades no caso de amostragem sem reposição.

Podese provar que $f(x)$ acima definida integra a 1 (não é muito fácil) .

Densidade Poisson

Esta densidade é usada principalmente para modelar o número de ocorrências de um evento "raro" (de probabilidade baixa) durante um intervalo de tempo especificado. Por exemplo, o número de acidentes numa estrada durante um fim de semana, o número de bactérias presentes numa solução após um certo período são, entre outros, eventos modelados pela distribuição de Poisson. Além disso, a distribuição de Poisson surge como um caso limite da distribuição $\text{Bin}(n, p)$ quando n é grande e p é pequeno (próximo de zero) e, neste contexto, é muito útil em aproximações numéricas.

v) Se X é uma variável aleatória contínua, sua função de distribuição é contínua. Se X é discreta, $F(x)$ é uma função contínua à direita, isto é, a função de distribuição apresenta "pulos" (descontinuidades) que só são "sentidos" quando nos aproximamos do ponto onde existe o "pulo" pela esquerda.

Relação entre a função densidade e a função de distribuição

Considere uma variável aleatória contínua com densidade $f(x)$ e função de distribuição $F(x)$. Então:

$$P(X \leq b) = \int_{-\infty}^b f(x) dx \therefore$$

$$\int_{-\infty}^b f(x) dx = F(x)$$

$$\frac{d}{dx} F(x) = f(x)$$

Ou seja, a densidade é a derivada da função de distribuição.

Distribuições discretas

Distribuição binomial

A densidade Binomial é uma das mais importantes em teoria da probabilidade. Ela surge como uma idealização matemática de diversas situações comuns na "vida real" e está intimamente ligada à amostragem com reposição. A situação clássica em que usamos uma densidade Binomial é a seguinte:

- Uma experiência tem apenas 2 resultados possíveis : "sucesso" e "falha", onde a probabilidade de "sucesso" é p e a probabilidade de "falha" é $q = 1 - p$.
- A experiência é repetida um número fixo (n) de vezes, sempre nas mesmas condições, de tal forma que as probabilidades de "sucesso" (p) e "falha" ($q = 1 - p$) se mantêm inalteradas a cada repetição. As diversas repetições da experiência são feitas de maneira independente, ou seja, o resultado de uma repetição não afeta o resultado das outras.
- A variável aleatória X que mede o número de "sucessos" nas n repetições da experiência é uma variável discreta, com valores possíveis $0, 1, 2, \dots, n$. Dizemos que esta variável tem densidade Binomial com parâmetros n e p , e escrevemos que $X \sim \text{Bin}(n, p)$.

Ou seja, esta última propriedade nos permite calcular a probabilidade de qualquer evento envolvendo a variável aleatória X . Para qualquer evento A definido no espaço da variável aleatória X , a probabilidade de A é apenas o somatório de todos aqueles valores de x que compõem A .

$f(x)$ é chamada densidade de probabilidade da variável aleatória X e dizemos que X é uma variável aleatória discreta.

Definição (Variável aleatória contínua, densidade de probabilidade contínua)

Seja X uma variável aleatória com espaço A a tal que:

i) $f(x) \geq 0 \forall x \in \mathbb{R}$

ii) $f(x)$ tem no máximo um número finito de descontinuidades em qualquer subintervalo finito de λ .

iii) Seja $A \subseteq \lambda$. A probabilidade de $X \in A$ é:

$$P(X \in A) = P(A) = \int_A f(x) dx$$

A variável aleatória X é uma variável aleatória contínua e $f(x)$ é sua densidade de probabilidade.

Seja X uma variável aleatória qualquer (contínua ou discreta). A função de distribuição da variável aleatória X é definida como:

$$F(x) = P(X \leq x)$$

Seja $f(x)$ a densidade de probabilidade de X . Então:

$$F(x) = \sum_{a \leq x} f(a)$$

Se X é variável aleatória discreta, ou

$$F(x) = \int_{-\infty}^x f(x) dx$$

Se X é variável aleatória contínua.

Propriedades da Função de Distribuição

i) $0 \leq F(x) \leq 1$ pois $0 \leq \Pr(X \leq x) \leq 1$.

ii) $F(x)$ é uma função não decrescente.

iii) $\lim_{x \rightarrow -\infty} F(x) = 0$

iv) $\lim_{x \rightarrow \infty} F(x) = 1$

Assim, a variável aleatória X é uma função que "transporta" a probabilidade de um espaço amostral S para um espaço λ de números reais.

Exemplo

Jogamos uma moeda duas vezes e estamos interessados no número de "caras" observado. O espaço amostral é:

$$S = \{c: \text{onde } c = \text{CaCa}, \text{CaCo}, \text{CoCa}, \text{CoCo}\}.$$

Podemos definir uma variável aleatória X como:

$$X(c) = \begin{cases} 0 & \text{se } c = \text{CaCa} \\ 1 & \text{se } c = \text{CaCo} \text{ ou } \text{CoCa} \\ 2 & \text{se } c = \text{CoCo} \end{cases}$$

Ou seja, X é o número de "caras" nas duas jogadas. O espaço da variável aleatória X é $\lambda = \{0, 1, 2\}$, um subconjunto dos inteiros.

Seja $A = \{x \in \lambda: x = 1\}$. Como definir a probabilidade do evento A ? É só olhar para o subconjunto S do espaço amostral cujos elementos c são tais que $X(c) \in A$, ou seja, $X(c) = 1$ aqui.

Neste caso, $S = \{\text{CaCo}, \text{CoCa}\}$, o evento "uma cara em duas jogadas", pois

$X\{\text{CaCo}\} = 1$ e $X\{\text{CoCa}\} = 1$. Assim

$$P\{A\} = P\{X \in A\} = P\{S\} = P\{X=1\}$$

Definição (Variável Aleatória Discreta, densidade de probabilidade discreta)

Seja X uma variável aleatória cujo espaço é o conjunto unidimensional λ . Dizemos que X é uma variável aleatória discreta se o número de valores possíveis de X é finito ou contável.

Por exemplo, se os valores possíveis de X são $\{0, 1, 2, \dots\}$ ou $\{2, 4, 6, 8, \dots\}$, ou ainda $\{1/3, 1/4, 1/5, \dots, 1/n, \dots\}$, X é uma variável aleatória discreta. A partir da definição de variável aleatória discreta podemos definir a densidade de probabilidade discreta, que é uma função da variável aleatória X que nos permite calcular probabilidades para todos os valores de X .

Seja $f(x)$ uma função tal que:

$$\text{i) } f(x) \geq 0 \forall x \in \lambda$$

$$\text{ii) } \sum_{x \in \lambda} f(x) = 1$$

$$\text{iii) } \forall A \subseteq \lambda, \Pr(A) = \Pr(X \in A) = \sum_{x \in A} f(x)$$

Seja S o espaço amostral e X uma função que "pega" elementos deste espaço (resultados da experiência) e os leva num subconjunto de números reais. Esta função X é chamada de variável aleatória.

Definição (Variável Aleatória)

Considere uma experiência aleatória com espaço amostral S . Seja c um elemento de S . Uma variável aleatória X é uma função que associa um único número real $X(c) = x$ a cada elemento do espaço amostral S . O espaço de X é o conjunto de números reais $\mathbb{R} = \{x : x = X(c) \forall c \in S\}$.

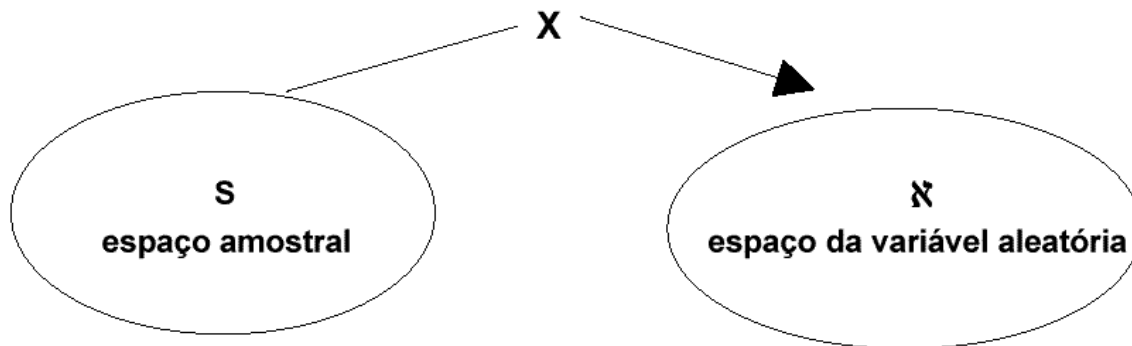


Figura 4- Espaço amostral espaço da variável aleatória

Seja X uma variável aleatória definida num espaço amostral S e seja λ o espaço de X .

Seja A um subconjunto de λ e s um subconjunto de S .

Já definimos a probabilidade de um evento $s \in S$, e agora gostaríamos de estender esta definição e falar da probabilidade de um evento $A \in \lambda$. Ou seja, o nosso objetivo agora é definir probabilidades a partir de valores possíveis da variável aleatória, sem referência explícita aos pontos do espaço amostral que deram origem àqueles valores da variável aleatória.

Na verdade, o estudo de probabilidades começa a se aprofundar com a definição de variável aleatória, que será o objeto principal das nossas atenções a partir de agora. A noção de variável aleatória é tão importante que freqüentemente nem nos preocupamos com o que está "por trás" delas, isto é, na prática muitas vezes ignoramos o espaço amostral.

Como definir $P(X \in A)$?

A maneira mais natural de fazer isso é associar a probabilidade do evento $X \in A$ à probabilidade do evento S no espaço amostral S .

Ou seja, se $A \in \lambda$, definimos:

$$P(X \in A) = P(S) \text{ onde } S = \{c \in S : X(c) \in A\}$$

A unidade de medida da variância é o quadrado da unidade de medida das observações. Assim, se os dados estão em metros, a variância é expressa em metros quadrados. Isso dificulta a interpretação da variância amostral. Para evitar isso trabalhamos com o desvio padrão amostral, definido a seguir.

-Desvio Padrão amostral

O desvio padrão amostral, denotado por s , é definido como a raiz quadrada positiva da variância amostral. Pelos comentários acima concluímos que s é sempre expresso nas mesmas unidades de medida que as observações na amostra. O desvio padrão da população é definido como a raiz quadrada da variância da população, e denotado por σ . Logo,

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$$

-Coeficiente de variação amostral

O coeficiente de variação amostral é definido como:

$$CV\% = \frac{s}{\bar{X}} \cdot 100$$

Onde s é o desvio padrão amostral e \bar{X} é a média amostral. A definição do coeficiente de variação para a população é análoga, substituindo-se σ por s e \bar{X} por μ .

O coeficiente de variação é útil quando comparamos a variabilidade de amostras com magnitudes ou unidades muito diferentes.

-Amplitude Amostral (range)

É a diferença entre o máximo e o mínimo da amostra, isto é :

$$R = X_n - X_1$$

VARIÁVEIS ALEATÓRIAS, FUNÇÕES DENSIDADE DE PROBABILIDADE

Seja S o espaço amostral. Muitas vezes este espaço não consiste num conjunto de números. Por exemplo, se jogamos uma moeda duas vezes, o espaço amostral é $S = \{\text{CaCa}, \text{CaCo}, \text{CoCa}, \text{CoCo}\}^1$, onde cada resultado tem a mesma probabilidade. No entanto, é difícil lidar com espaços amostrais desta maneira, é muito mais fácil trabalhar com quantidades numéricas. Neste exemplo específico poderíamos estar interessados no número de "caras" nas 2 jogadas, e seria interessante definir uma função que associasse um número a cada resultado no espaço amostral.

¹ Onde Ca indica "cara", Co indica "coroa".

Por exemplo, nos dois gráficos a seguir as populações têm a mesma média (μ), mas certamente a segunda distribuição tem maior dispersão.

Figura 3- Distribuição 1

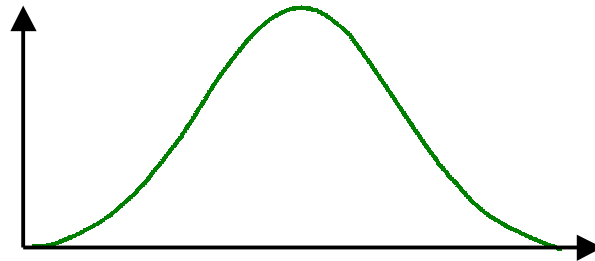
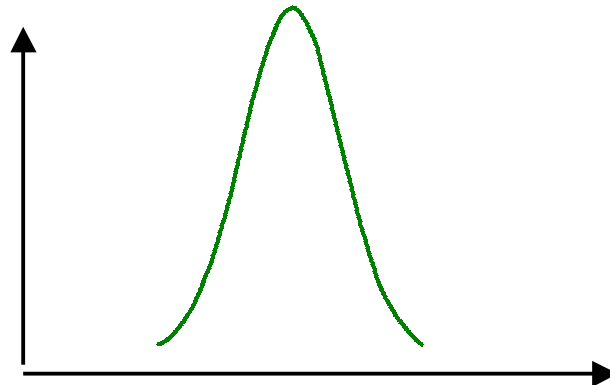


Figura 3- Distribuição 2



-Variância amostral

É a medida mais comum de dispersão . A variância amostral, denotada por s^2 é definida como:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

Onde \bar{X} é a média amostral, já definida.

O análogo da variância amostral para a população será denotado por σ^2 .

Nota: A variância (da amostra ou da população) é sempre maior ou igual a zero.

Eventualmente mostraremos como a média amostral \bar{X} pode ser usada para estimar ("chutar") a média da população (μ).

Considere agora a sua amostra x_1, x_2, \dots, x_n e suponha que você ordene a amostra, de tal forma que x_1 é o menor elemento da amostra, x_2 é o segundo menor elemento, ..., x_n é o maior elemento da amostra. Os valores x_1, x_2, \dots, x_n são chamados de estatísticas de ordem da amostra. Outras medidas de tendência central e de dispersão serão definidas a partir das estatísticas de ordem.

-Mediana

A mediana amostral é definida a partir das estatísticas de ordem como:

$m = X_{\frac{n+1}{2}}$ se n , o tamanho da amostra é ímpar ou,

$m = \frac{X_{\frac{n}{2}} + X_{\frac{n+1}{2}}}{2}$ se n é par.

Por exemplo, se existem 10 observações na amostra, a mediana equivale à média entre x_5, x_6 . Se a amostra contém 11 elementos, a mediana é x_5 .

Analogamente ao caso da média, também podemos definir uma mediana para a população.

A mediana amostral tem uma vantagem sobre a média amostral : ela é menos influenciada por observações extremas do que a média amostral.

Por exemplo, suponha que os dados na amostra são : 1, 3, 4, 2, 7, 6, 8. A média amostral é 4.43, e a mediana é 4. Se os dados agora são : 1, 3, 4, 2, 7, 2519, 8, a média amostral é 363.43, mas a mediana continua sendo 4.

É claro que este exemplo é radical, mas ilustra bem o fato da mediana ser mais "robusta" ao encontrar observações discrepantes do resto da amostra.

-Moda

A moda amostral é simplesmente a observação mais freqüente na amostra. Se os meus dados são : 1, 4, 8, 12, 5, 4, 4, 7, a moda é 4, o valor que ocorreu mais vezes. Também é possível definir a moda de uma população, como veremos mais tarde.

Medidas de Dispersão

As medidas de tendência central não são as únicas medidas necessárias para caracterizar uma amostra (ou população). Precisamos também saber o quanto as observações na amostra estão "espalhadas".

| | | |
|------------------|-----|-------------------|
| 0-50 KWh | 127 | 127/1122 = 11.3 % |
| 51-100 KWh | 199 | 199/1122 = 17.7 % |
| 101-150 KWh | 225 | 20.1 % |
| 151-300 KWh | 384 | 34.2 % |
| acima de 300 KWh | 187 | 16.7 % |

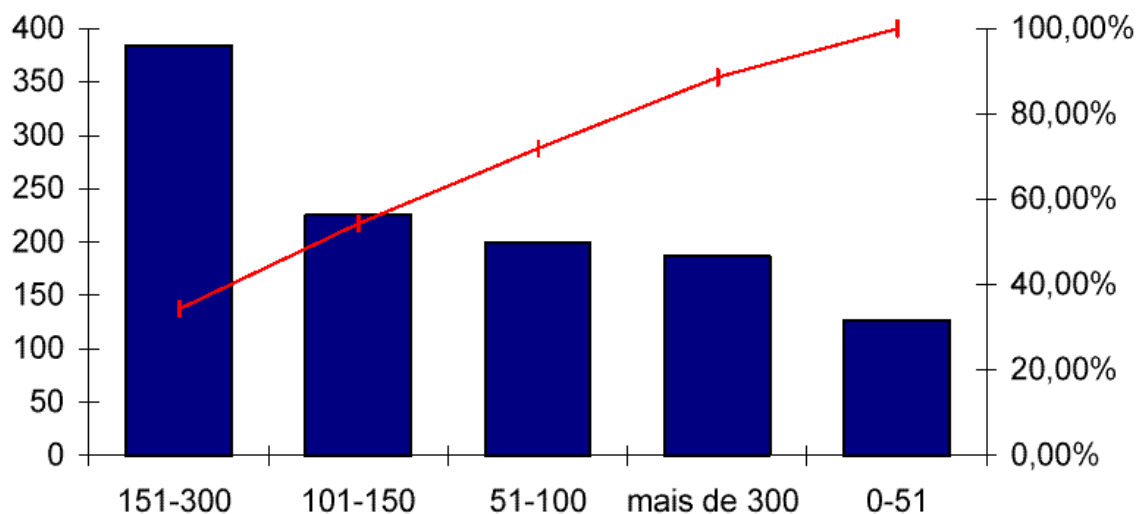


Figura 2- Diagrama de Pareto do número de domicílios por faixa de consumo.

A classe de consumo predominante na amostra é a classe 151-300 kWh mensais. O eixo do lado direito do gráfico exibe as frequências relativas acumuladas.

Os diagramas de Pareto são muito úteis em controle de qualidade, pois nos permitem ver que tipo de defeitos são os mais frequentes.

Medidas numéricas

A partir de agora suponha que os dados observados na amostra são x_1, x_2, \dots, x_n . Note que n é o tamanho da amostra. A partir dos x 's vamos encontrar números que resumem as características da amostra. Vamos estar interessados em 2 tipos principais de medidas numéricas: as que caracterizam a localização do centro da amostra e as que caracterizam a dispersão dos dados.

Medidas de Localização ou de tendência central

-Média Amostral

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

A média amostral é uma medida que indica onde está o "centro" da sua amostra.

Entretanto, o "centro" da população também pode ser definido, como veremos mais tarde, e esta medida será chamada de média da população, e denotada por μ .

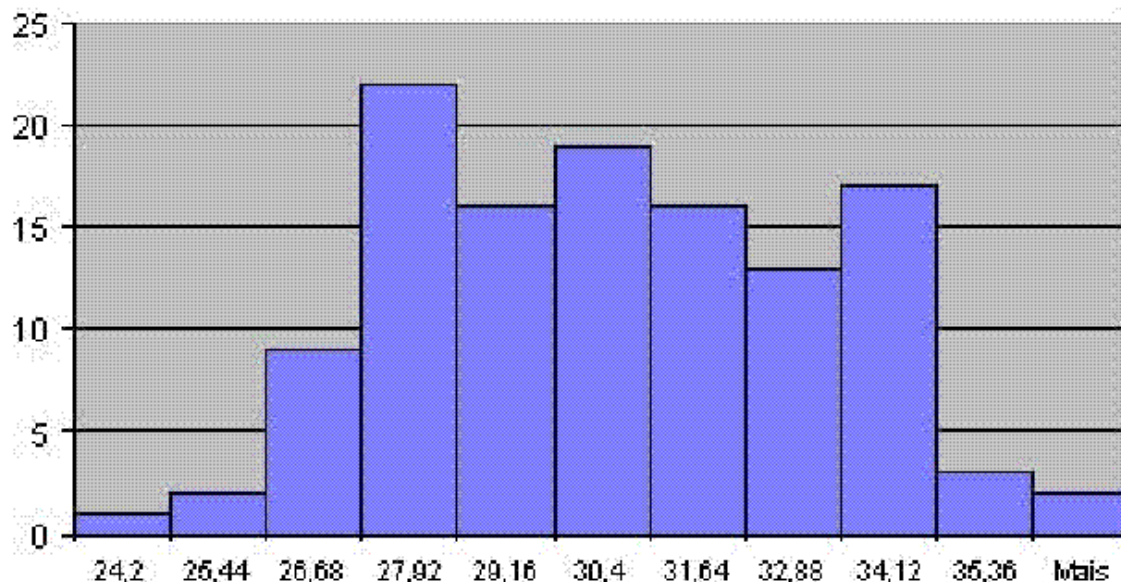


Figura 2- Distribuição de temperaturas

Diagrama de Pareto

Um diagrama de Pareto é também um gráfico de barras, usado para dados inteiros, tais como o número de objetos que apresentam diversos tipos de defeitos ou o número de ocorrências de algum evento de interesse.

Como fazer um diagrama de Pareto?

1) Faça um gráfico de barras colocando a frequência de cada tipo de evento no eixo vertical, e arranjando os eventos em ordem decrescente de ocorrência.

Assim, a primeira barra corresponde ao evento que ocorre com mais frequência, a segunda barra diz respeito ao segundo evento mais frequente, e assim por diante.

2) Crie um eixo vertical no lado direito do seu gráfico contendo as frequências relativas acumuladas. Faça uma linha juntando as frequências relativas acumuladas e a superponha ao gráfico de barras.

Exemplo

Os dados a seguir representam a distribuição de domicílios residenciais por classe de consumo de energia elétrica na área de concessão da Light. Os dados referem-se a uma pesquisa realizada em dezembro de 1995. O tamanho da amostra é 1122 domicílios.

Tabela 3- Distribuição do número de domicílios por faixa de consumo

| Faixas de consumo | Número de domicílios | Frequência relativa |
|-------------------|----------------------|---------------------|
|-------------------|----------------------|---------------------|

puramente prática, pois este número nos permitiu encontrar intervalos de classe de comprimento 1.9 em todas as classes, exceto a primeira, e todas as classes terminam com uma temperatura que é um número inteiro e par. Pura conveniência!

Tabela 2- Tabela de frequências -dados de temperatura

| Classe | Frequência | Frequência Relativa | Freq. Relativa Acumulada |
|---------------|-------------------|----------------------------|---------------------------------|
| 24-26 graus | 7 | $7/120 = 5.83 \%$ | 5.83 % |
| 26-28 graus | 31 | $31/120 = 25.83 \%$ | 31.66 % |
| 28-30 graus | 26 | $26/120 = 21.67 \%$ | 53.33 % |
| 30-32 graus | 26 | $26/120 = 21.67 \%$ | 75.00 % |
| 32-34 graus | 25 | $25/120 = 20.83 \%$ | 95.83 % |
| 34-36 graus | 3 | $3/120 = 2.50 \%$ | 98.33 % |
| 36-38 graus | 2 | $2/120 = 1.67 \%$ | 100 % |

A Tabela de frequências já nos permite responder a diversas outras questões. Por exemplo, a grande maioria (69.17 %) das temperaturas máximas está entre 26.1 e 32 graus. Também percebemos que temperaturas máximas acima de 34.1 graus são incomuns (apenas 5 dentre as 120).

Veja que outras conclusões você consegue obter a partir deste diagrama. A partir de uma Tabela de frequências podemos facilmente construir um histograma.

Histograma

Histograma é um gráfico de barras, onde o eixo vertical contém as frequências (ou frequências relativas) e o eixo horizontal contém os intervalos de classes. Muitas vezes faz-se a área de cada barra igual a frequência relativa de cada classe, de tal forma que a área total sob o histograma é 1 (100 %). O histograma a seguir foi produzido automaticamente pelo Excel. Você pode verificar que os pontos médios dos intervalos são diferentes dos que especificamos no diagrama de frequência.

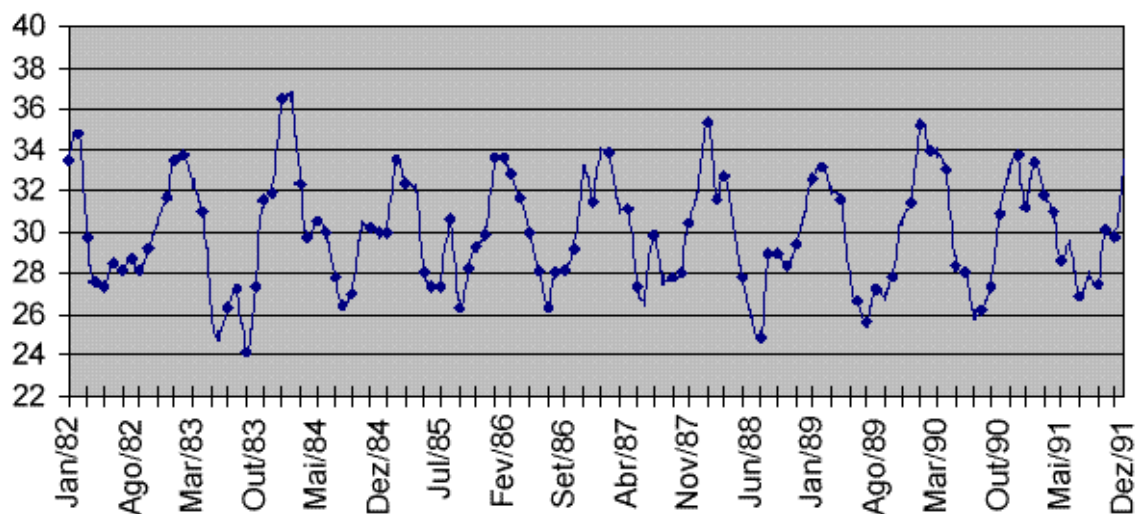


Figura 1- Temperaturas máximas (1982 a 1991)

O primeiro passo é fazer a distribuição de frequência dos seus dados. Isto é simplesmente uma medida mais compacta de representação dos dados. Você divide as temperaturas em intervalos (chamados intervalos de classe) e conta quantas observações caem em cada intervalo. Bem simples, né?

A escolha do número de intervalos é meio arbitrária. O importante é garantir que o número de classes não seja nem muito grande nem muito pequeno. Se o número de classes for muito pequeno, fica difícil verificar as diferenças entre as classes. Ao contrário, se o número de classes for muito grande, existirão muito poucas observações em cada classe.

O primeiro passo é ordenar os dados (se for possível fazê-lo automaticamente, senão, não vale a pena). Isto torna um pouco mais fácil a colocação dos dados em cada classe. Neste caso eu decidi considerar 7 classes para as temperaturas. A primeira vai de 24 a 26 graus, a segunda vai de 26.1 a 28 graus e assim sucessivamente. O diagrama de frequências encontrado está a seguir.

Nota: Escolha do número de classes num diagrama de frequência Seja n o número de intervalos num diagrama de frequência. Recomendase escolher n entre 5 e 20.

Quanto maior o número de observações, maior o número de intervalos.

Geralmente usase n igual à raiz quadrada do número total de observações, o que neste caso daria $\sqrt{120} \approx 11$. Para facilitar a visualização normalmente usamos intervalos com o mesmo comprimento (ou quase). Também muitas vezes o primeiro intervalo é descrito como "abaixo de um certo valor" e o último como "acima de um certo valor". Neste exemplo usamos $n = 7$, por uma questão

| | | | | | | | | | |
|--------|-------|--------|-------|--------|-------|--------|-------|--------|-------|
| jun-82 | 28.50 | jun-83 | 24.98 | jun-84 | 30.00 | jun-88 | 25.80 | jun-90 | 28.00 |
| jul-82 | 28.20 | jul-83 | 26.30 | jul-84 | 27.80 | jul-88 | 24.80 | jul-90 | 26.00 |
| ago-82 | 28.70 | ago-83 | 27.20 | ago-84 | 26.40 | ago-88 | 29.00 | ago-90 | 26.20 |
| set-82 | 28.10 | set-83 | 24.20 | set-84 | 27.00 | set-88 | 28.90 | set-90 | 27.40 |
| out-82 | 29.20 | out-83 | 27.40 | out-84 | 30.30 | out-88 | 28.40 | out-90 | 30.90 |
| nov-82 | 30.53 | nov-83 | 31.60 | nov-84 | 30.20 | nov-88 | 29.40 | nov-90 | 33.10 |
| dez-82 | 31.67 | dez-83 | 31.90 | dez-84 | 30.00 | dez-88 | 31.20 | dez-90 | 33.70 |
| jan-85 | 30.00 | jan-86 | 33.60 | jan-87 | 33.80 | jan-89 | 32.60 | jan-91 | 31.20 |
| fev-85 | 33.50 | fev-86 | 33.60 | fev-87 | 33.90 | fev-89 | 33.20 | fev-91 | 33.40 |
| mar-85 | 32.40 | mar-86 | 32.80 | mar-87 | 31.10 | mar-89 | 32.00 | mar-91 | 31.80 |
| abr-85 | 32.10 | abr-86 | 31.70 | abr-87 | 31.10 | abr-89 | 31.60 | abr-91 | 31.00 |
| mai-85 | 28.00 | mai-86 | 30.00 | mai-87 | 27.30 | mai-89 | 27.70 | mai-91 | 28.60 |
| jun-85 | 27.30 | jun-86 | 28.20 | jun-87 | 26.70 | jun-89 | 26.70 | jun-91 | 29.40 |
| jul-85 | 27.30 | jul-86 | 26.30 | jul-87 | 29.90 | jul-89 | 25.70 | jul-91 | 26.90 |
| ago-85 | 30.70 | ago-86 | 28.00 | ago-87 | 27.70 | ago-89 | 27.20 | ago-91 | 27.90 |
| set-85 | 26.30 | set-86 | 28.10 | set-87 | 27.85 | set-89 | 26.90 | set-91 | 27.50 |
| out-85 | 28.30 | out-86 | 29.20 | out-87 | 28.00 | out-89 | 27.80 | out-91 | 30.10 |
| nov-85 | 29.90 | nov-86 | 33.10 | nov-87 | 30.40 | nov-89 | 30.50 | nov-91 | 29.80 |
| dez-85 | 29.90 | dez-86 | 31.40 | dez-87 | 32.10 | dez-89 | 31.50 | dez-91 | 33.30 |

O gráfico apresentado na Figura 1é muito útil, mas certamente ele não conta à estória toda.... Por exemplo, qual será a temperatura média de todos os meses? Dentre os 120 meses, em quantos a temperatura média esteve entre 28 e 33 graus? Qual o percentual de temperaturas entre 22 e 25 graus? Tomando-se os 120 pontos, quais os valores de temperatura tais que 90 % dos meses têm temperaturas entre estes dois valores?

Podemos pensar nestas, e numa infinidade de outras questões. O fato é que um simples gráfico da temperatura versus o tempo não fornece as respostas.

Medidas Numéricas:

- média amostral,
- mediana amostral,
- desvio padrão amostral,
- variância amostral.

Suponha que você já definiu :

- quais as características de interesse da população.
- qual o tamanho da amostra
- você já coletou os dados
- você já entrou com os dados no computador e verificou (e corrigiu) possíveis erros de digitação.

E agora?????

A primeira coisa a fazer é tentar "bolar" um gráfico. Cá entre nós, é meio difícil tentar chegar a alguma conclusão olhando para uns 200 números diferentes. Daí o comentário sobre um gráfico valer tanto quanto mil palavras!

Exemplo

A próxima tabela nos dá a média das temperaturas máximas mensais na estação Santa Cruz no período entre Janeiro de 1982 e Dezembro de 1991. O que fazer com todos estes 120 números?

A coisa mais sensata é fazer um gráfico da temperatura versus o índice de tempo (mês e ano). Este gráfico vai revelar o óbvio, isto é, que as temperaturas no verão são mais altas que no inverno! Além disso, a gente vai perceber que existe um comportamento sazonal nos dados, ou seja, dentro de cada ano a evolução da temperatura se repete mais ou menos da mesma maneira. O gráfico também nos dá uma idéia do quanto a temperatura está variando em todo o período. Por exemplo, vamos verificar que a temperatura máxima nestes 10 anos está sempre acima de 22 graus.

Tabela 1- Temperatura máxima (média das máximas) na estação de Santa Cruz (RJ) .

| Mês | Ano | Mês | Ano | Mês | Ano | Mês | Ano | Mês | Ano |
|--------|-------|--------|-------|--------|-------|--------|-------|--------|-------|
| jan-82 | 33.55 | jan-83 | 33.51 | jan-84 | 36.50 | jan-88 | 35.30 | jan-90 | 35.20 |
| fev-82 | 34.80 | fev-83 | 33.69 | fev-84 | 36.60 | fev-88 | 31.60 | fev-90 | 34.00 |
| mar-82 | 29.80 | mar-83 | 32.42 | mar-84 | 32.40 | mar-88 | 32.70 | mar-90 | 33.80 |
| abr-82 | 27.60 | abr-83 | 31.00 | abr-84 | 29.70 | abr-88 | 30.40 | abr-90 | 33.00 |
| mai-82 | 27.40 | mai-83 | 25.81 | mai-84 | 30.50 | mai-88 | 27.80 | mai-90 | 28.40 |

4- dentro os bebedores de Brahma, quantas cervejas eles tomam por semana e a que classe social eles pertencem ? Existe alguma relação entre estas

4- variáveis (consumo e classe social) ?

Como você pode ver, existe uma infinidade de características de interesse que podem ser medidas numa amostra.

Em resumo:

A partir de uma amostra coletamos informações que nos permitirão aprender alguma coisa interessante sobre a população.

ESTATÍSTICA DESCRITIVA

(" A picture is worth one thousand words")

Em termos de custos, estatística é uma coisa super eficiente, pois muitas vezes (como nos exemplos) a população é enorme, e não é viável medir as características de interesse em cada elemento da população. Na verdade, pode-se provar que, para populações muito grandes, uma amostra de cerca de 600 ou 1000 "indivíduos" fornece resultados bastante confiáveis sobre as características da população.

Suponha agora que você obteve uma amostra e, dentro desta amostra você coletou dados numéricos (por exemplo, o % de audiência da TV Globo nos domingos à noite).

O que fazer com isso? Existem duas possibilidades:

-você pode simplesmente descrever estes dados numéricos através de gráficos e tabelas. Isto é chamado de estatística descritiva. A maioria das pesquisas de mercado faz só isso, que é sem dúvida, muito importante.

-você pode tentar tirar conclusões sobre as características da população a partir dos dados observados na amostra. Isto se chama estatística inferencial (ou simplesmente estatística!), e será a nossa grande preocupação neste curso.

Para que a gente consiga fazer isso, é necessário ter uma noção bastante abrangente de Probabilidades, e isto irá ocupar grande parte do nosso curso. Na verdade a estatística descritiva surgiu muito antes da estatística inferencial. Esta última depende da especificação de modelos matemáticos baseados numa noção fundamental, que é a de "probabilidade".

As ferramentas usuais da estatística descritiva são as seguintes:

Gráficos:

- histograma,
- diagramas de Pareto.

POPULAÇÃO E AMOSTRA

Por que estatística é importante?

Porque nos permite entender e lidar com a noção de variabilidade.

Um exemplo típico é: produção de parafusos. Uma fábrica produz parafusos, que devem ter seu diâmetro dentro de certas especificações. Ao medirmos o diâmetro de 100 parafusos produzidos ao acaso existirão variações individuais. Estas variações são importantes? Até que ponto as variações observadas são aceitáveis? Em geral um número em Estatística não é apenas um número! A ele associamos uma medida de incerteza ou variabilidade.

População = coleção de todos os elementos cujas características desejamos conhecer.

Os elementos (ou "indivíduos") na população não são necessariamente pessoas!

Amostra = subconjunto da população cujas características serão medidas. A amostra será usada para descobrir características da população.

Exemplos:

1) População: eleitores na cidade do Rio de Janeiro

Amostra: 650 eleitores escolhidos aleatoriamente (ao acaso)

Característica de interesse : percentual de eleitores que planejam votar num candidato X nas próximas eleições.

2) População: automóveis Uno Mille produzidos em 1995

Amostra: todos os automóveis produzidos em agosto de 1995

Característica de interesse: número de defeitos apresentados nos primeiros 3 meses de uso, quilometragem média , e uma possível relação entre estas duas variáveis.

3) População: todos os domicílios com TV na cidade do Rio de Janeiro.

Amostra: 1000 domicílios com TV escolhidos ao acaso.

Característica de interesse: percentual de audiência de cada emissora de TV a cada dia da semana no horário de 18 às 22 horas.

4) População: população acima de 15 anos na cidade do Rio de Janeiro.

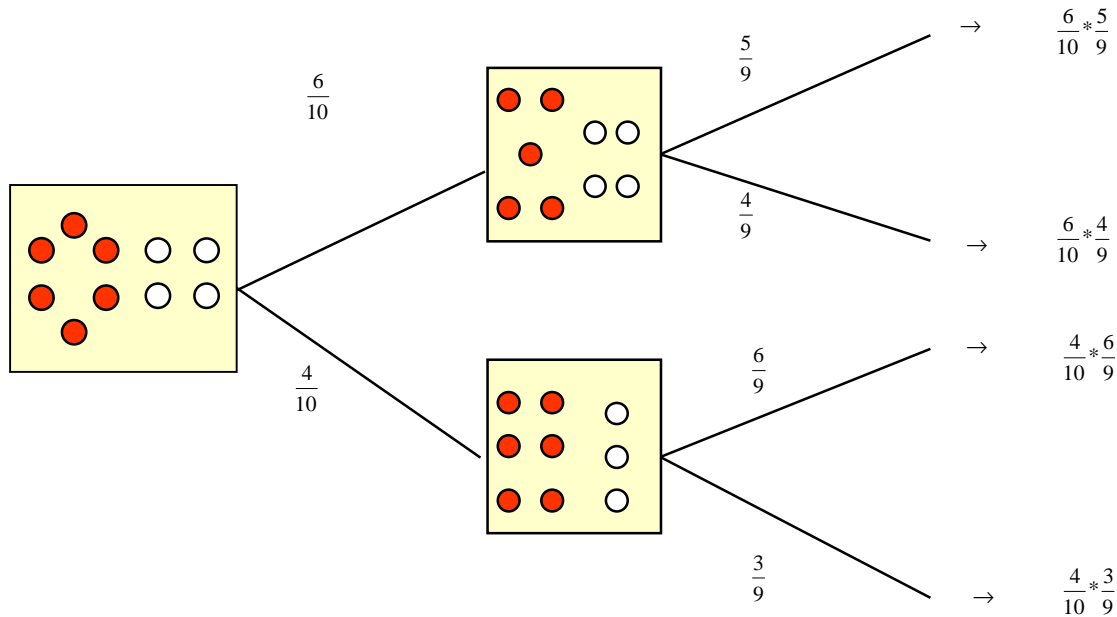
Amostra: 200 pessoas com mais de 15 anos.

Características de interesse =

1-percentual de bebedores de cerveja

2-dentre os bebedores de cerveja, quantos são homens ?

3-dentre os bebedores de cerveja, quantos preferem Brahma ?



Independência

Dois eventos A e B são independentes se $P\{A / B\} = P\{A\}$ ou $P\{B / A\} = P\{B\}$.

No exemplo anterior $P\{A/B\} = \frac{5}{9} \neq \frac{6}{10} = P\{A\}$, ou seja, A e B não são independentes.

Se o sorteio da 2ª bola for com reposição:

$P\{A / B\} = P\{\text{sortear 1 bola vermelha dentre 6 vermelhas e 4 brancas}\}$

$$= \frac{6}{10} = P\{A\} \text{ então os eventos são independentes.}$$

Regra do produto para eventos independentes

Se A e B são eventos independentes,

$$P\{A \cap B\} = P\{B\} \cdot P\{A / B\} = P\{B\} \cdot P\{A\}$$

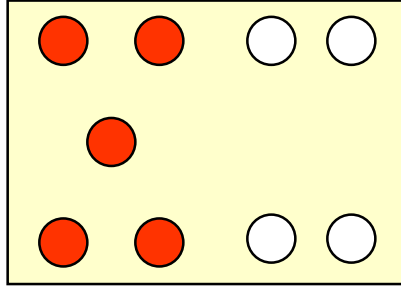
Exemplo

Considerando o exemplo anterior, qual a probabilidade de serem retiradas 2 bolas vermelhas (no sorteio com reposição)?

Calculando a probabilidade de A , temos:

$$P\{A \cap B\} = P\{A\} \cdot P\{B\}$$

$$= \frac{6}{10} \cdot \frac{6}{10} \\ = 0,36$$



Se B ocorreu, isto é, saiu vermelha na primeira retirada, então

$P\{A/B\} = P\{\text{sortear 1 bola vermelha dentre 5 vermelhas e 4 brancas}\}$

$$P\{A/B\} = \frac{5}{9}$$

$P\{A / B^c\} = P(\text{sortear 1 bola vermelha dentre 6 vermelhas e 3 brancas}) = 6/10$

Portanto:

$$P\{A\} = P\{B\} \cdot P\{A / B\} + P\{\bar{B}\} \cdot P\{A / \bar{B}\}$$

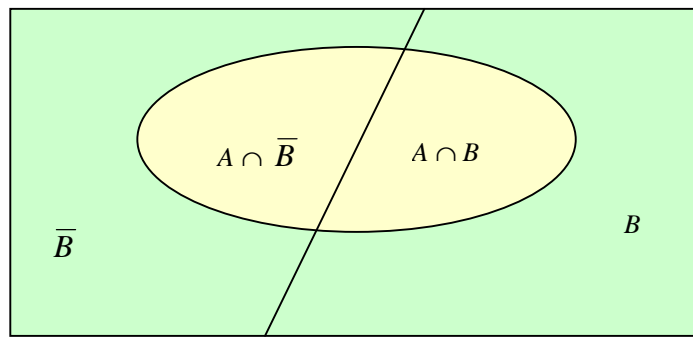
$$P\{A\} = \frac{5}{9} \cdot \frac{6}{10} + \frac{6}{9} \cdot \frac{4}{10}$$

$$P\{A\} = \frac{6}{10} \left(\frac{5}{9} + \frac{4}{9} \right)$$

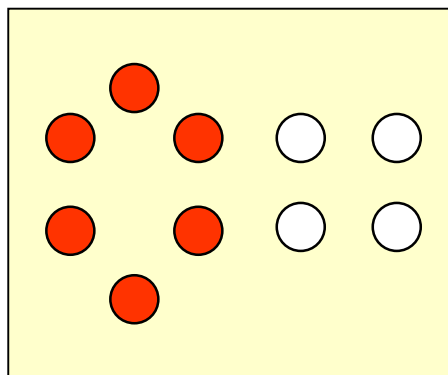
$$P\{A\} = \frac{6}{10}$$

Podemos fazer o diagrama em árvore ou árvore de probabilidades da situação descrita neste exercício.

$$P\{A\} = \frac{6}{10} \cdot \frac{5}{9} + \frac{4}{10} \cdot \frac{6}{9} = \frac{6}{10}$$

**Exemplo**

Em uma urna, há 10 bolas: 4 brancas e 6 vermelhas. Duas bolas são sorteadas sucessivamente, SEM reposição. Qual é a probabilidade da 2ª bola ser vermelha?



Se A é a probabilidade da 2ª bola sorteada ser vermelha, queremos calcular $P\{A\}$. Seja B a probabilidade da 1ª bola sorteada ser vermelha.

$$P\{B\} = \frac{6}{10}$$

$$P\{\bar{B}\} = 1 - P(B)$$

$$= 1 - \frac{6}{10}$$

$$= \frac{4}{10}$$

Note que:

$$P\{L / M\} = \frac{\frac{\text{nº jovens do sexo masculino e que sabem ler}}{\text{nº total de jovens}}}{\frac{\text{nº jovens do sexo masculino}}{\text{nº total de jovens}}}$$

$$P\{L / M\} = \frac{P\{M \cap L\}}{P\{M\}} \quad (*)$$

Seja A e B eventos de um experimento aleatório qualquer. Imitando (*), podemos dizer que a probabilidade condicional de A dado B (nota-se por $P(A / B)$) é definida como:

$$P\{A / B\} = \frac{P\{A \cap B\}}{P\{B\}} \quad (**)$$

Por exemplo, a probabilidade de ser do sexo masculino dado que lê é dada por:

$$P\{M / L\} = \frac{P\{M \cap L\}}{P\{L\}} = \frac{0,388}{0,843} = 0,460$$

Regra do produto

Da equação (**) obtemos a **regra do produto** para a probabilidade da interseção de dois conjuntos:

$$P\{A \cap B\} = P\{A / B\} \cdot P\{B\} \quad (**)$$

válida para quaisquer eventos A e B de S .

Regra da probabilidade total

Sejam A e B dois eventos.

Há duas maneiras de A ocorrer: ou A e B ocorrem $\{A \cap B\}$ ou A e \bar{B} ocorrem $\{A \cap \bar{B}\}$.

Deste modo, $A = \{A \cap B\} \cup \{A \cap \bar{B}\}$, onde $A \cap B$ e $A \cap \bar{B}$ são conjuntos disjuntos.

Pela regra da soma $P\{A\} = P\{A \cap B\} + P\{A \cap \bar{B}\}$.

Pela regra do produto $P\{A\} = P\{B\} \cdot P\{A / B\} + P\{\bar{B}\} \cdot P\{A / \bar{B}\}$ (regra da probabilidade total).

$$P\{M\} = \frac{\text{nº de jovens do sexo masculino de } S}{\text{nº de jovens de } S} = \frac{48.245}{101.850} = 0,473$$

$$F = \overline{M} \Rightarrow P\{F\} = P\{\overline{M}\} = 1 - P\{M\} = 1 - 0,473 = 0,527$$

$$P\{M/L\} = \frac{\text{nº de jovens do sexo masculino e que sabem ler de } S}{\text{nº total de jovens (S)}} = \frac{39.557}{101.850}$$

$$\begin{aligned} P\{M \cup L\} &= P\{M\} + P\{L\} - P\{M \cap L\} \\ &= 0,473 + 0,843 - 0,388 \\ &= 0,928 \end{aligned}$$

Relembrando a interpretação da probabilidade, seja A um evento de um experimento aleatório de um espaço amostral. Consideramos duas formas de se atribuir probabilidades aos eventos de um espaço amostral:

- $P\{A\}$ é uma crença (subjetiva) que se deposita na ocorrência de A .
- Interpretação freqüencista (objetiva)

$$f_n(A) = \frac{\text{nº de repetições que } A \text{ ocorre}}{n}$$

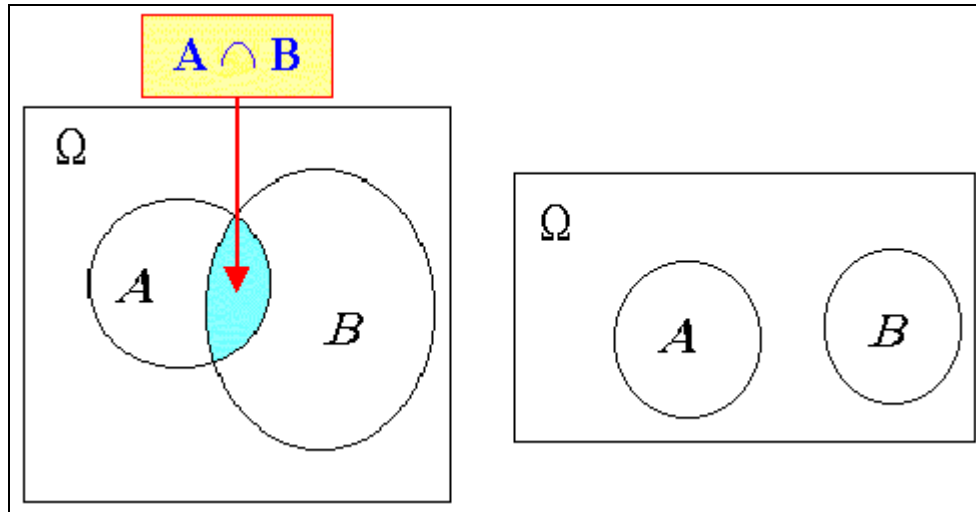
Quando n cresce: $f_n(A) \approx P(A)$, isto é, quando n cresce a probabilidade é aproximada pelo valor da freqüência relativa.

No exemplo anterior, se soubermos que o jovem sorteado é do sexo masculino, qual é a probabilidade de que saiba ler? Temos uma *informação parcial*: o jovem é do sexo masculino.

Vamos designar a probabilidade de L quando se sabe que o jovem é do sexo masculino por $P(L/M)$ e denominá-la probabilidade (condicional) de L dado M .

É natural atribuímos:

$$\begin{aligned} P\{L/M\} &= \frac{\text{nº de jovens que sabem ler dentre aqueles do sexo masculino}}{\text{nº total de jovens do sexo masculino}} \\ &= \frac{39.577}{48.249} = 0,820 \end{aligned}$$



Exemplo

Dados do Censo Demográfico de 91 publicado pelo IBGE relativos aos habitantes de Sergipe, na faixa etária entre 20 a 24 anos com relação às variáveis Sexo e Leitura.

| Sexo | Lê | Não lê | Total |
|------------------|--------|--------|---------|
| Masculino | 39.577 | 8.672 | 48.249 |
| Feminino | 46.304 | 7.297 | 53.601 |
| Total | 85.881 | 15.969 | 101.850 |

Um jovem entre 20 e 24 anos é escolhido ao acaso em Sergipe.

- Ω = conjunto de 101.850 jovens de Sergipe, com idade entre 20 e 24 anos.

Eventos de interesse:

- M = jovem sorteado é do sexo masculino = jovens do sexo masculino de Ω
- F = jovem sorteado é do sexo feminino
- L = jovem sorteado sabe ler
- $M \cap L$ = jovem sorteado é do sexo masculino e sabe ler
- $M \cup L$ = jovem sorteado é do sexo masculino *ou* sabe ler

Podemos obter algumas probabilidades:

$$P\{L\} = \frac{\text{nº de jovens que sabem ler de } S}{\text{nº de jovens de } S} = \frac{85.881}{101.850} = 0,843$$

- \emptyset é o evento impossível

Operações com eventos

Dados dois eventos A e B de um mesmo espaço amostral:

- $A \cap B$ é o evento em que A e B ocorrem simultaneamente
- $A \cup B$ é o evento em que A ocorre ou B ocorre (ou ambos)
- \bar{A} é o evento em que A não ocorre.

Exemplo

Considerando o lançamento de um de um dado e os eventos A , B , C , D e E definidos há pouco, temos:

- $B \cap D = \text{sair uma face par e maior que 3}$
- $B \cap D = \{2, 4, 6\} \cap \{4, 5, 6\} = \{4, 6\}$
- $B \cap C = \text{sair uma face par e ímpar}$
- $B \cap C = \{2, 4, 6\} \cap \{1, 3, 5\} = \emptyset$ (B e C são disjuntos)
- $C = \bar{B}$
- $B = \bar{C}$
- $B \cup D = \text{sair uma face par ou maior que 3}$
- $B \cup D = \{2, 4, 6\} \cup \{4, 5, 6\} = \{2, 4, 5, 6\}$
- $B \cup C = \text{sair uma face par ou ímpar}$
- $B \cup C = \{2, 4, 6\} \cup \{1, 3, 5\} = \{1, 2, 3, 4, 5, 6\}$

Probabilidade

Definição

A probabilidade é uma função que atribui um número aos eventos de Ω (se A é um evento de Ω , $P(A)$ é a probabilidade de A), que satisfaz:

1. $0 \leq P\{A\} \leq 1$
2. $P\{\emptyset\} = 0$, $P\{\Omega\} = 1$
3. **Regra da soma:** dados dois eventos mutuamente exclusivos A e B de Ω ,

$$P\{A \cup B\} = P\{A\} + P\{B\}$$
4. Regra de soma para eventos quaisquer:

$$P\{\bar{A}\} = 1 - P\{A\}$$
 para todo evento A .

$$\frac{3}{6} = \frac{\text{número de casos favoráveis}}{\text{número de casos possíveis}}$$

Modelos matemáticos para os experimentos aleatórios

Definição de espaço amostral

O espaço amostral é denotado por Ω , o conjunto de todos os resultados possíveis de um experimento.

Exemplo

$S = \{\text{chove, não chove}\}$

Em geral, temos interesse em eventos particulares do experimento.

- **evento A:** chove
- $A = \{\text{chove}\} \subset S$
- $\{\text{chove}\} \rightarrow$ subconjunto de S

Exemplo

$S = \{1, 2, 3, 4, 5, 6\}$

- **evento B:** sair face par
- $B = \{\text{sair face par}\} = \{2, 4, 6\} \subset S$
- $\{\text{sair face par}\} = \{2, 4, 6\} \rightarrow$ subconjunto de S

Exemplo

Ainda considerando $S = \{1, 2, 3, 4, 5, 6\}$, podemos definir outros eventos tais como:

- **evento C:** sair uma face *ímpar*
- $C = \{1, 3, 5\}$
- **evento D:** sair uma face maior que 3
- $D = \{4, 5, 6\}$
- **evento E:** sair face 1
- $E = \{1\}$

Resumo

A um experimento aleatório está associado a um espaço amostral S .

O evento A ocorre se o resultado do experimento pertence a A .

Os conjuntos S e \emptyset também são eventos:

- S é o evento certo.

INTRODUÇÃO À TEORIA DA PROBABILIDADE

O conceito de experimento

Designaremos por experimento todo processo que nos fornece dados:

- pode ser a observação de um experimento natural: observação astronômica, meteorológica, sísmica;
- observação de um experimento controlado para testar a fadiga de materiais, verificar o resultado de um exame de sangue, etc;
- pesquisa de opinião para saber: quantos estudantes fumam na Universidade; e
- quantos eleitores tem intenção de votar num candidato A em uma eleição.

Nos experimentos mencionados pode-se notar que a incerteza sempre esta presente, o que quer dizer que se estes experimentos forem repetidos em idênticas condições não se pode determinar qual o resultado que ocorrerá.

A incerteza esta associada a chance de ocorrência que atribuímos ao resultado de interesse.

Exemplo

Vai chover no litoral no fim de semana?

- **Conjunto de possibilidades:** $S = \{\text{chove, não chove}\}$

Para calcular a probabilidade de chover podemos, ou usar a intuição (subjetivo) ou usar a frequência relativa dos últimos dez fins de semana em que choveu (objetivo).

Exemplo

Lançamento de dado: você ganha se sair uma **face par**.

- **Conjunto de possibilidades:** $S = \{1, 2, 3, 4, 5, 6\}$
- **Conjunto de possibilidades favoráveis:** $\{2, 4, 6\}$
- **Probabilidade de vitória** = ?

Supondo que um dado é equilibrado, é natural atribuir a probabilidade:

- 2- Para determinar o número médio de pessoas que vivem numa casa, foi planejado visitar 1000 casas. Como o entrevistador não achou ninguém em 133 das 1000 casas que deveria visitar, visitou outras, até completar 1000. Este procedimento está correto. Por que?
- 3- Um fabricante que conhecer características das donas-de- casa que usam sabão em pó que fabrica. Foram feitas duas amostras: Na primeira se perguntava à dona de casa se ela usava o sabão em pó do fabricante e não outra os entrevistadores iam de casa em casa, e pediam para ver o sabão que estava sendo usado. As duas amostras devem dar resultados diferentes. Por que?
- 4- Os alunos do curso colegial estão distribuídos em três séries:1^a, 2^a e 3^a. Pretende-se fazer um estudo para conhecer os hábitos de fumar desses alunos. Decidiu-se então fazer uma entrevista com 150 alunos. Para construir a amostra foram feitas duas propostas. A primeira sugeria escrever os nomes de todos os alunos em cédulas, misturar bem e tirar 150. A Segunda sugeria pedir aos professores que escolhessem os alunos mais representativos de cada série. Discuta essas duas propostas e , se for o caso, faça uma terceira

- Amostras sistemáticas
- Amostras estratificadas
- Amostras de conveniência

A amostra casual simples

Os elementos são escolhidos para formar a amostra por processo casual ou aleatório, ou seja por sorteio. Exemplo: Imagine os alunos de uma escola. Como fazemos para obter uma amostra casual simples? Basta, atribuir número para cada aluno. E sorteei.

Amostra sistemática

Os elementos são escolhidos para formar a amostra por critério estabelecido a priori pelo pesquisador. Imagine as casa de um bairro. Como obter uma amostra sistemática? Atribua um número para cada casa. Toda a casa, cujo o número terminar em determinado dígito (por exemplo) pertencerá a amostra.

Amostra estratificada

Obtemos amostra estratificada quando a população se apresenta dividida em Estratos isto é, em diferentes grupos. Imagine os empregados de uma indústria. Como procedemos uma amostra? A população está dividida em estratos – no caso – quanto a qualificação, diretoria, escritório e oficina. Tire uma amostra casual ou sistemática dentro de cada estrato. A amostra estratificada é formada por elementos provindo de diferentes estratos.

Observação final

A informação obtida com base em amostras só pode ser estendida para a população de onde a amostra proveio.

Exercícios

- 1- O que é amostragem?

- 3- Usamos amostras por economia, pois para fazer uma pesquisa é necessário, entrevistador, questionários, transporte, etc...
- 4- Usamos amostras para maior precisão – você tem a possibilidade de examinar (analisar) uma amostra com maior cuidado do que a população porque a amostra é menor.

A validade da amostra

A amostra é parte da população. Descrevemos o todo (a população) tendo examinado parte (amostra). Por exemplo, os resultados dos exames das amostras de sangue indicam o que o paciente tem. Imagine se fosse preciso tirar todo o sangue para chegar a um diagnóstico. Em uma linguagem técnica, chamamos, isto de inferência.

Muito cuidado com a inferência...Antes de fazer inferência pare e pense!!

- A amostra veio da população que você estava estudando? Ou é uma amostra tendenciosa??

A amostra é tendenciosa quando as pessoas que respondem são diferentes das pessoas que não respondem.

- Se você escolher um indivíduo de cada grupo, os indivíduos que pertencem a grupos grande têm menor probabilidade de serem escolhidos para a amostra.
- As pessoas podem ser influenciadas por determinadas formas de perguntas.

Cuidados básicos para coleta de dados

- 1- colete todos os dados relevantes – evite coleta de dados desnecessário
- 2- quando a pessoa não responder, escreva “não sabe”, ou “não declarou”, assim, você sabe que não esqueceu de perguntar
- 3- Estabeleça forma de registro de dados.

Técnicas de amostragem

De acordo com a técnica utilizada, temos:

- Amostras casuais simples,

- 3- Imagine que as taxas da inflação , nos últimos cinco meses foram 0,5%, 0,61%, 0,8% e 0,9%. Faça dois gráficos de linha: Um para mostrar grande aumento da inflação e outros para mostrar pequeno aumento da inflação.
- 4- Suponha que, em 5 dias consecutivos, choveu em determinado local. A quantidade de chuva, em milímetros, foi respectivamente: 1 1,5 2,5 3 3,5. Os dados colocados em gráfico, ficam praticamente sobre uma reta. No entanto, a meteorologia alega que não poder prever a quantidade de chuva que cairá daqui a 100 dias, com base na reta. Por que?

AMOSTRAGEM

Quando se fala em população, sempre se imagina o conjunto de habitantes de um país, uma cidade, uma região. Para um pesquisador, no entanto, o termo população tem sentido bem mais geral.

População é o conjunto de elementos que têm, em comum, determinada característica. Por exemplo, falamos em “população” dos empregados de uma indústria, isto é, em todas as pessoas que têm, em comum, a característica de serem empregados dessa indústria.

Para estudar dados de toda a população são feitos censos. No Brasil, se fazem censos demográficos nos anos terminados em zero.

No entanto, nem sempre se faz um censo. Muitas vezes se faz uma amostra.

Amostra é todo conjunto não vazio e com número menor de elementos do que a população a qual foi extraída.

Você pode estar se perguntando, Por que usamos amostras??

- 1- Usamos amostra por que existem populações infinitas. Por exemplo, quantas vezes podemos lançar um dados?? Por maior que seja o número imaginado, sempre se pode lançar o dado uma outra vez. Então todo conjunto de valores observado é uma amostra.
- 2- Usamos amostras quando a população é tão grande que, para fins práticos, podemos admitir como infinita.

$$\hat{y} = -1,1 + 1,7 x \text{ e assim, para } x=2 \Rightarrow \hat{y} = 2,3$$

$$\text{para } x=4 \Rightarrow \hat{y}=5,7$$

4- Como temos dois pontos, podemos traçar uma reta

Observação

- Dado um valor de X, que não foi observado na mostra, você pode prever Y.

Lembre-se

| x | y | x | y |
|---|---|-----|---|
| 1 | 1 | 1 | 1 |
| 2 | 2 | 2 | 2 |
| 3 | 4 | 3 | 4 |
| 4 | 5 | 4 | 5 |
| 5 | 8 | 4,5 | ? |
| | | 5 | 8 |

- Use a reta de regressão

$$Y = -1,1 + 1,7 X$$

Então, para $x=4,5$

$$y = -1,1 + 1,7 * 4,5 \Rightarrow y = 6,55$$

Importante!!!! Evite estimar valores de Y fora do intervalo estudado de X

A previsão de y é melhor, quanto os pontos estiverem bem próximos a reta.

Exercícios

- 1- Obtenha as retas de regressão para os dados apresentados no exercício 2 e 3 do tópico anterior.
- 2- Se os filhos fossem exatamente 5 cm mais altos do que os respectivos pais, como seria a reta de regressão que daria a estatura dos filhos em função da estatura dos pais?

$$Y = a + b X, \text{ em que}$$

- a: coeficiente linear, é a altura em que a reta corta o eixo dos Y
- b: coeficiente angular, dá a inclinação da reta

O cálculo de **a** e de **b**, são:

$$b = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} \quad \bar{y} = \frac{\sum y}{n} \quad \bar{x} = \frac{\sum x}{n}$$

$$a = \bar{y} - b\bar{x}$$

Para calcular **a** e **b**, faz-se da seguinte forma:

| | | | | | |
|---|---|---|---|---|---|
| X | 1 | 2 | 3 | 4 | 5 |
| Y | 1 | 2 | 4 | 5 | 8 |

1- Primeiro calcule xy , x^2 e os somatórios

| x | y | x^2 | xy |
|----|----|-------|----|
| 1 | 1 | 1 | 1 |
| 2 | 2 | 4 | 4 |
| 3 | 4 | 9 | 12 |
| 4 | 5 | 16 | 20 |
| 5 | 8 | 25 | 40 |
| 15 | 20 | 55 | 77 |

← Somatórios

2- Agora, basta substituir na fórmula

$$b = \frac{77 - \frac{(15 \cdot 20)}{5}}{55 - \frac{(15)^2}{5}} = \frac{17}{10} = 1,7 \Rightarrow \boxed{b = 1,7}$$

$$\bar{y} = \frac{20}{5} = 4 \quad \bar{x} = \frac{15}{5} = 3$$

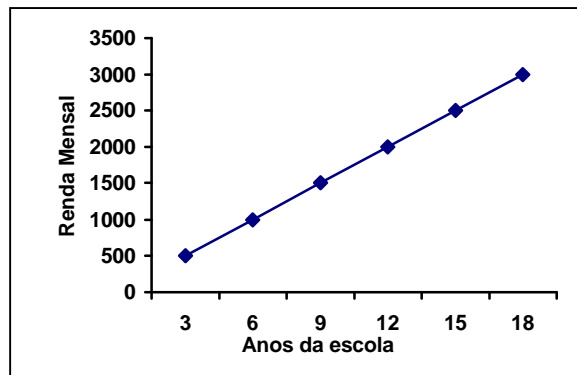
$$a = \bar{y} - 1,7 \cdot \bar{x} = 4 - (1,7 \cdot 3) = 4 - (5,1) = -1,1 \Rightarrow \boxed{a = -1,1}$$

3- Basta, colocar no gráfico

Ache dois pontos da reta...

| Entrevistados | Anos na Escola | Renda Mensal (R\$) |
|---------------|----------------|--------------------|
| F | 3 | 500,00 |
| B | 6 | 1.000,00 |
| C | 9 | 1.500,00 |
| E | 12 | 2.000,00 |
| D | 15 | 2.500,00 |
| A | 18 | 3.000,00 |

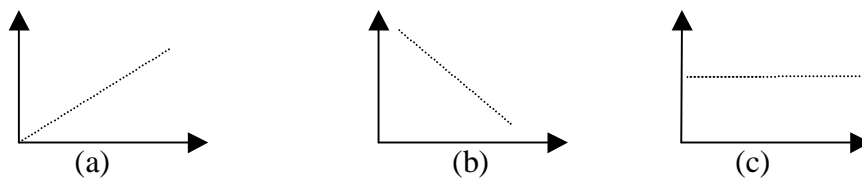
Para estudar uma função de variáveis traçamos o gráfico de linhas, em que a



variável independente (X), está no eixo das abcissas, e a variável Y está no eixo das ordenadas.

Observações

1- O aspecto do gráfico de linhas permite algumas conclusões.



A figura (a) ilustra uma situação em que está havendo um aumento na variável observada, já a figura (b), mostra um declínio, contudo na figura (c) nada se pode concluir.

Os pontos traçados no diagrama de dispersão podem ficar praticamente sobre uma linha reta. E assim podemos então traçar a reta que mostra Y em função de X.

A equação da reta é:

| | | | | | |
|----------------|---|---|---|---|----|
| \overline{Y} | 2 | 4 | 6 | 8 | 10 |
|----------------|---|---|---|---|----|

b)

| | | | | | |
|---|---|---|---|---|---|
| X | 1 | 2 | 3 | 4 | 5 |
| Y | 2 | 4 | 6 | 8 | 0 |

3- Calcule os coeficientes de correlação para os dados dos dois conjuntos a seguir.

Depois faça os diagramas de dispersão

| | | | | | |
|---|---|---|---|---|---|
| X | 1 | 2 | 3 | 4 | 5 |
| Y | 2 | 3 | 2 | 3 | 2 |

| | | | | | |
|---|---|---|---|---|---|
| X | 1 | 2 | 3 | 4 | 5 |
| Y | 1 | 2 | 3 | 2 | 1 |

4- Um estudo mostrou que, na Inglaterra, a taxa de morte por doenças do coração era maior entre motoristas de ônibus do que entre cobradores. A princípio se pensou que o tipo de trabalho fosse maior causa da doença, mas depois se notou que o tamanho dos uniformes, que forneciam aos motoristas, era sistematicamente maior que o tamanho dos uniformes que forneciam aos cobradores. O que isto sugere a você?

FUNÇÃO DE VARIÁVEIS

Observando nosso cotidiano, pode-se observar que algumas variáveis variam em função de outras variáveis. Algumas são muito conhecidas, por exemplo: Estatura é função da idade, peso é a função de estatura, a produção é função da quantidade de fertilizante usada em uma lavoura, renda mensal em função de quantos anos de estudo a pessoa tem., e muitos outros.

Se Y varia em função de X dizemos que:

X: variável independente

Y: variável dependente.

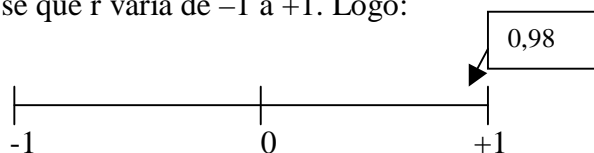
Um exemplo: Imaginem um estudo em que se procure estabelecer a correlação entre X = anos completos de frequência à escola, e Y = renda mensal. Seja uma amostra de seis entrevistados.

$$r = \frac{\sum xy - \frac{\sum x \cdot \sum y}{n}}{\sqrt{\left[\sum x^2 - \frac{(\sum x)^2}{n} \right] \left[\sum y^2 - \frac{(\sum y)^2}{n} \right]}}$$

$$r = \frac{77 - \frac{15 \cdot 20}{5}}{\sqrt{\left[55 - \frac{(15)^2}{5} \right] \left[110 - \frac{(20)^2}{5} \right]}} = 0,98$$

3- Interprete o resultado:

Lembre-se que r varia de -1 a $+1$. Logo:



$r = 0,98$ é um valor muito alto. Isto significa que existe alta correlação positiva entre X e Y

Cuidado com a interpretação

Suponha que um estudante, após uma aula de correlação, observou em um jornal que, o número de internações por desidratação estava aumentando, e no mesmo jornal, havia uma manchete dizendo que, o número de refrigerantes vendidos na mesma cidade estava aumentando, também. O estudante pode concluir que os refrigerantes causam desidratação????? **NÃO**. Quando a temperatura aumenta, aumentam a venda de refrigerantes e os casos de desidratação.

É preciso avaliar, o que realmente está influenciando uma determinada situação, já que muitas conclusões não são fidedignas. Portanto, correlação não implica em causa e efeito.

Exercícios

- 1- O que é correlação?
- 2- Calcule o coeficiente de correlação para os dados dos dois conjuntos abaixo. Faça os diagramas de dispersão. Discuta por que os valores de r são tão diferentes, embora os dados sejam semelhante?

a)

| | | | | | |
|---|---|---|---|---|---|
| X | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|

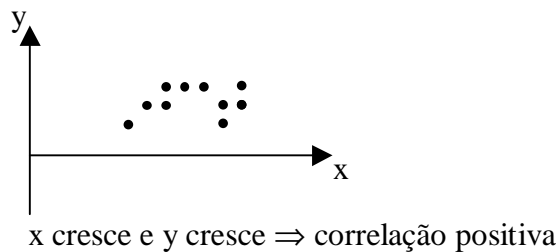
Correlação linear simples

Quando duas variáveis crescem no mesmo sentido dizemos que entre elas existe uma correlação positiva.

Quando duas variáveis crescem em sentidos opostos dizemos que entre elas existe correlação negativa.

Quando uma variável cresce e a outra varia ao acaso dizemos que entre elas existe correlação nula.

Para melhor visualizar, pode-se dispor os pontos (dados) em um gráfico, chamado de diagrama de dispersão. Abaixo, tem-se um exemplo de correlação positiva.



Exemplo:

| | | | | | |
|---|---|---|---|---|---|
| X | 1 | 2 | 3 | 4 | 5 |
| Y | 1 | 2 | 4 | 5 | 8 |

Para calcular o **r**:

1- Ordene a tabela

| x | y | x^2 | y^2 | xy |
|----|----|-------|-------|----|
| 1 | 1 | 1 | 1 | 1 |
| 2 | 2 | 4 | 4 | 4 |
| 3 | 4 | 9 | 16 | 12 |
| 4 | 5 | 16 | 25 | 20 |
| 5 | 8 | 25 | 64 | 40 |
| 15 | 20 | 55 | 110 | 77 |

2- Substitua os totais na fórmula

CORRELAÇÃO

O objetivo do estudo da correlação é a determinação da força do relacionamento entre duas observação emparelhadas. O termo correlação significa literalmente, co-relacionamento, pois indica até que ponto os valores de uma variável estão relacionados com os de outra. Há muitos casos em que pode existir um relacionamento entre duas variáveis. Consideremos, por exemplo, questões como estas:

- a idade e a resistência física estão correlacionadas?
- Pessoas de maior renda tendem a apresentar melhor escolaridade?
- O sucesso num emprego poder ser esperado com base no resultado de testes?
- A temperatura parece influenciar a taxa de criminalidade?
- Estudantes com maior capacidade de leitura tendem a obter melhores resultados em matemática?

Problemas como esses se prestam à análise de correlação. O resultado de tal análise é o coeficiente de correlação – um valor que quantifica o grau de correlação. O coeficiente de correlação é representado por **r**.

Este coeficiente possui uma característica, ele varia entre -1 e 1 .

A formula de **r** é a seguinte:

$$r = \frac{\sum xy - \frac{\sum x \cdot \sum y}{n}}{\sqrt{\left[\sum x^2 - \frac{(\sum x)^2}{n} \right] \left[\sum y^2 - \frac{(\sum y)^2}{n} \right]}}$$

Para calcular o valor de **r** é preciso calcular:

- $\sum xy$: soma dos produtos xy
- $\sum x$: soma dos valores de x
- $\sum y$: soma dos valores de y
- $\sum x^2$: soma dos quadrados de x
- $\sum y^2$: soma dos quadrados de y
- $(\sum x)^2$: quadrado da soma de x
- $(\sum y)^2$: quadrado da soma de y

| A | B | C | |
|-----|-----|-----|----------------------|
| 100 | 600 | 500 | |
| 0 | 400 | 500 | |
| 100 | 600 | 500 | |
| 0 | 400 | 500 | |
| 500 | 500 | 500 | <i>Médias iguais</i> |

Exercícios

1- Calcule a amplitude, a variância e o desvio padrão dos seguintes conjuntos

- a) 1 2 3 4 5
- b) 2 4 6 8 10
- c) 2 4 6 8
- d) 2 4 6 8 0

2- Sem calcular qualquer medida de dispersão, assinale o conjunto que apresenta a maior dispersão e o conjunto que apresenta a menor dispersão

- a) 2 4 2 4 2 4
- b) 0 6 0 6 0 6
- c) 3 3 3 3 3 3

3- Para obter o peso médio de um conjunto de ratos tanto se pode pesar cada leitão por vez como pesar todos os leitões de uma vez. No entanto, obtemos mais informações se pesarmos um rato por vez? Por que?

4- De acordo com alguns, estatística é uma ciência engraçada porque, se nós formos ao restaurante e eu comer um frango, enquanto você não come nenhum, em média comemos cada um, meio frango. Consequentemente, ambos deveríamos estar satisfeitos. Discuta a validade dessa média.

Portanto, podemos concluir que o mês de junho os preços variaram mais!

Desvio Padrão

O desvio padrão é a raiz quadrada da variância. Representa-se por s .

$$s = \sqrt{s^2}$$

Exemplo:

Num conjunto de dados a média é $\bar{x} = 20\text{m}$, e a variância é $s^2 = 225 \text{ m}^2$.

Quanto vale o desvio padrão?

$$s = \sqrt{225} = 15$$

ou seja, o desvio padrão é de 15m.

Observação:

1- Conjuntos com médias iguais podem ser diferentes

| A | B | C | |
|------|-----|-----|-----------------|
| 1000 | 600 | 500 | |
| 0 | 400 | 500 | |
| 1000 | 600 | 500 | |
| 0 | 400 | 500 | |
| 500 | 500 | 500 | ← Médias iguais |

Então, lembre-se não olhe apenas a média, verifique a variâncias.

2- Variabilidade e dispersão são sinônimos. Então tanto se fala de medidas de dispersão, como em medidas de variabilidade.

3- A média representa tanto melhor o conjunto de dados quanto menos dor a variância.

Um exemplo: Observe os dados dos conjuntos A, B e C. A média representa melhor C do que B, e B melhor do que A.

VARIÂNCIA DE DADOS DISPOSTOS EM TABELAS DE DISTRIBUIÇÃO DE FREQUÊNCIAS

Diante das tabelas abaixo, como avaliar, qual que apresenta a maior variação?

Número de itens vendidos, segundo
segundo

O preço em reais em maio

| Preço | Nº de itens |
|-------|-------------|
| 10 | 4 |
| 11 | 4 |
| 13 | 10 |
| 14 | 2 |

Número de itens vendidos,

o preço em reais em junho

| Preço | Nº de itens |
|-------|-------------|
| 14 | 4 |
| 15 | 10 |
| 19 | 10 |
| 20 | 1 |

Fórmula da variância

$$s^2 = \frac{\sum x^2 f - \frac{(\sum xf)^2}{n}}{n - 1}$$

Então para o mês de maio

| Preço | Nº de itens | x^2 | $x^2 f$ | $x f$ |
|-------|-------------|-------|---------|-------|
| 10 | 4 | 100 | 400 | 40 |
| 11 | 4 | 121 | 484 | 44 |
| 13 | 10 | 169 | 1690 | 130 |
| 14 | 2 | 196 | 392 | 28 |
| Total | 20 | - | 2966 | 242 |

Logo,

$$s^2 = \frac{2966 - \frac{(242)^2}{20}}{19} \cong 1,989$$

A variância de junho é feita da mesma forma e seu cálculo é $s^2 \cong 5,073$

| | |
|-----------------|--------------------|
| 6 | 36 |
| $\Sigma x = 20$ | $\Sigma x^2 = 104$ |

2- Substitua os resultados na fórmula

$$s^2 = \frac{104 - (20)^2}{4 - 1}$$

$$s^2 = \frac{104 - 100}{3} = \frac{4}{3}$$

$$s^2 = 1,33$$

Lembre-se a variância mede a variabilidade

Observações:

- 1- Se os dados não variam, a variância é obrigatoriamente zero
- 2- Quanto maior for a variabilidade dos dados, maior será a variância
- 3- A unidade de medida de variância é igual ao quadrado da unidade de medida dos dados (porque os valores são elevados ao quadrado)

Lembra-se dos casos dos alunos, observe como a variância mede a variabilidade

| Aluno | Notas | | | | Variância |
|--------|-------|----|----|---|---------------|
| João | 5 | 5 | 5 | 5 | $s^2 = 0,00$ |
| Paulo | 4 | 6 | 4 | 6 | $s^2 = 1,33$ |
| Pedro | 0 | 10 | 10 | 0 | $s^2 = 16,67$ |
| Júnior | 10 | 5 | 5 | 0 | $s^2 = 33,33$ |

- As notas de João não variam: a variância é zero
- As notas de Paulo variaram menos do que as notas de Pedro. Observe na tabela: variância das notas de Paulo é menor do que a variância das notas de Pedro.
- As notas de Júnior variaram mais do que as notas de todos os outros. Veja na tabela: a variância das notas de Júnior é a maior.

Desvio ou afastamento

Desvio ou afastamento é a diferença entre determinado valor e a média.

Variância

Para medir a variabilidade os estatísticos usam a variância.

Exemplo: São dadas as notas de um aluno em quatro provas

4 6 4 6

1- Vamos indicar as notas pela letra “x” e escrever as notas em coluna;

| X |
|---|
| 4 |
| 6 |
| 4 |
| 6 |

3- A variância é indicada por s^2 e dada pela fórmula:

$$s^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}, \text{ em que}$$

$\sum x^2$: soma dos quadrados

$(\sum x)^2$: quadrado da soma

1- Organize o cálculo:

| x | x^2 |
|---|-------|
| 4 | 16 |
| 6 | 36 |
| 4 | 16 |

Júnior e Pedro, para que pudessem observar as notas. João dizia que sempre tirava 5,0, Paulo comentou que as suas notas variavam, e por fim, Pedro, disse que com ele, era “Tudo ou Nada”.

As notas dos três alunos, seguem abaixo

| AVISO | | | | | |
|--------|---|-------|----|----|---|
| Nome |  | Notas | | | |
| João | | 5 | 5 | 5 | 5 |
| Paulo | | 4 | 6 | 4 | 6 |
| Pedro | | 0 | 10 | 10 | 0 |
| Júnior | | 10 | 5 | 5 | 0 |

Vamos calcular a amplitude das notas

| Aluno | Notas | | | | Amplitude |
|--------|-------|----|----|---|-----------|
| João | 5 | 5 | 5 | 5 | $5-5=0$ |
| Paulo | 4 | 6 | 4 | 6 | $6-4=2$ |
| Pedro | 0 | 10 | 10 | 0 | $10-0=10$ |
| Júnior | 10 | 5 | 5 | 0 | $10-0=10$ |

Observação: A amplitude igual a zero, significa que não houve variabilidade
Quanto maior a amplitude, maior a variabilidade.

Mas, uma coisa vocês podem estar se perguntando. Por que as notas de Pedro e Júnior têm a mesma amplitude?

Veja:

- 1- A amplitude se baseia apenas nos valores extremos (maior e menor).
- 2- Conjuntos diferentes de dados podem Ter a mesma amplitude
- 3- Então a amplitude não mede bem a variabilidade
- 4- Usa-se a amplitude apenas porque é fácil de calcular e de interpretar.

- 3- Dado um conjunto de número para obter a média, a mediana e a moda, qual dessas medidas corresponderá, necessariamente, a um valor numérico do conjunto?
- 4- As quatro pessoas que estão reunidas numa sala têm, em média 20 anos. Se uma pessoa com 40 anos entra na sala, qual passa a ser a idade média do grupo?
- 5- O preço médio de um produto era \$150,00. Numa liquidação, o produto foi oferecido por \$125,00. Qual foi o percentual de redução?

MEDIDAS DE VARIABILIDADE

Vamos imaginar o seguinte exemplo: Tem- se um comerciante, e um professor. Digamos que a renda média mensal das duas categorias, seja de R\$700,00. Será que essa informação indica que as duas distribuições de renda são necessariamente semelhantes? Muito pelo contrário, poderia descobrir que elas diferem – e muito – num outro aspecto importante, qual seja, o fato de as rendas dos professores concentrarem-se ao redor de R\$700,00, enquanto que as rendas do comerciante espalham-se mais, o que reflete, portanto, que em um determinado mês o comerciante vende mais, e em outros vende pouco.

Tal fato demonstra que necessitamos, além de uma medida de tendência central, de um índice que indique o grau de dispersão dos escores em torno do centro da distribuição (isto é em torno da média). Numa palavra, precisamos de uma medida indicadora do que costumeiramente se chama de variabilidade (também designada variação ou dispersão). Voltando a um exemplo, dado anteriormente, poderíamos dizer que a distribuição de renda do professor tem menor variabilidade do que a distribuição de renda do comerciante.

Estaremos tratando aqui, das medidas de variabilidade mais conhecidas: amplitude total, a variância, e o desvio padrão.

Amplitude

A amplitude é a diferença entre o maior e o menor valor do conjunto de dados.

Exemplo: Uma revista estudantil gostaria de saber qual a variabilidade das notas em uma determinada sala de aula. Assim, foram escolhidos três alunos, João, Paulo,

- De modo geral, a média possui certas propriedades matemáticas que a tornam atraente. Além disso, a ordenação dos dados para determinar a mediana pode ser cansativa, e o cálculo da mediana não pode ser feito com máquinas de calcular, ao contrário do que ocorre com a média;
- Comparada com a média e a mediana, a moda é a menos útil das medidas para problemas estatísticos, porque não se presta à análise matemática, ao contrário do que ocorre com as outras duas medidas.

| Medida | Vantagens | Limitações |
|---------|---|--|
| Média | - Reflete cada valor - Possui propriedades matemáticas atraentes | É influenciada por valores extremos |
| Mediana | Menos sensível a valores extremos do que a média | Difícil de determinar para grande quantidade de dados |
| Moda | Valor mais freqüente – maior quantidade de dados concentrados num determinado ponto | - Não se presta a análise matemática - Pode não ser moda para certos conjuntos de dados |

Exercícios

1- Determine a média, a mediana e a moda dos seguintes conjuntos de dados

- a) 8 3 0 6 8
- b) 8 16 2 8 6
- c) 4 16 10 6 20 10
- d) 0 -2 3 -1 5
- e) 2 -1 0 1 2 1 9

2- Imagine que você está dirigindo um carro numa estrada e observa que o número de carros que você ultrapassa é igual ao número de carros que ultrapassam você. Nesse caso a velocidade do seu carro corresponde a que medida de tendência central?

| | |
|------------|---------|
| 11 a 14 | 483.495 |
| 15 ou mais | 19.486 |

Exemplo: Encontra a moda

1- Dados

5 4 3 6 6 3 1 6 2

2- Ordene

1 2 3 3 4 5 6 6 6

3- A moda é: 6

1 2 3 3 4 5 6 6 6

Observação: Note que existem conjuntos com duas modas ou mais modas

Seqüência 1 5 5 7 8 9 10 10

Seqüência 2 1 1 2 6 9 10 10 19

E ainda, existem conjuntos sem moda

1 7 10 15 20

Comparação entre Média, Mediana e Moda

Há um momento em que o pesquisador procura uma medida de tendência central para a sua situação particular de pesquisa. E você usará qual medida?? A moda e média ou a mediana. Sua decisão envolve vários fatores, tais como:

- A média é uma medida que é influenciada por cada valor do conjunto, inclusive extremos;
- Por outro lado, a mediana é relativamente insensível aos valores extremos;

4- Calcule a média dos dois valores

$$\frac{62 + 64}{2} = 63$$

Observação: Quando existem número iguais, todos eles devem ser ordenados!!

Moda

Moda é o valor que ocorre com maior frequência. Um exemplo:

A distribuição de famílias, residentes em domicílios particulares, segundo o número de pessoas, no Brasil, em 1980, possuem as seguintes frequências.

| Número de pessoas | Número de famílias |
|-------------------|--------------------|
| 1 | 1.554.972 |
| 2 | 4.440.200 |
| 3 | 5.028.241 |
| 4 | 4.839.945 |
| 5 | 3.772.972 |
| 6 | 2.543.195 |
| 7 a 10 | 4.124.242 |
| 11 a 14 | 483.495 |
| 15 ou mais | 19.486 |

A moda neste caso, é 3 pessoas por família.

| Número de pessoas | Número de famílias |
|-------------------|--------------------|
| 1 | 1.554.972 |
| 2 | 4.440.200 |
| 3 | 5.028.241 |
| 4 | 4.839.945 |
| 5 | 3.772.972 |
| 6 | 2.543.195 |
| 7 a 10 | 4.124.242 |

1- Seqüência de dados:

71 82 57 68 78 75 64 61 85

2- Ordene

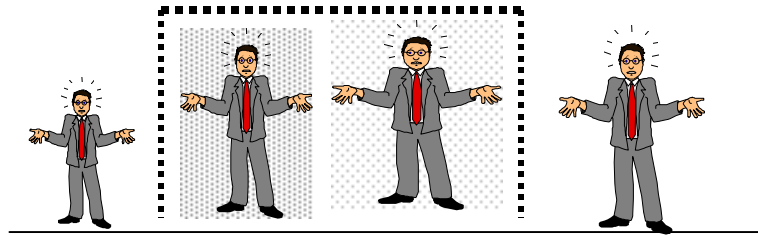
57 61 64 68 71 75 78 82 85

3- Mediana

57 61 64 68 **71** 75 78 82 85

Observação: metade dos dados são iguais ou menores do que a mediana.

Imagine um **número par** de indivíduos



A mediana será a média das estaturas dos dois indivíduos que ocupam a posição central.

Exemplo: Encontre a mediana

1- seqüência de dados

62 54 82 54 75 64

2- Ordene

54 54 62 64 75 82

3- Mediana

54 54 **62 64** 75 82

| Preço | Número de itens vendidos | Produto |
|-------|--------------------------|---------|
| 10 | 4 | 40 |
| 11 | 4 | 44 |
| 13 | 10 | 130 |
| 14 | 2 | 28 |
| Total | 20 | 242 |

4- Divida a soma pelo total de itens vendidos obtendo o preço médio

$$\text{Preço médio} = \frac{242}{20} = 12,1$$

Mediana

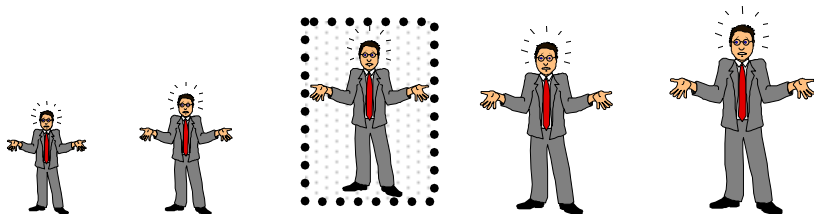
Mediana é o valor que ocupa a posição central dos dados ordenados. Imagine um número ímpar de elementos.



Para encontrar a estatura mediana, coloque os indivíduos em ordem crescente.



Mediana é a estatura do indivíduo que ocupa a posição central



Exemplo: Encontre a mediana.

A média representa o ponto de equilíbrio, é o valor em torno do qual os dados se distribuem.

Imagine a seguinte situação...

Um comerciante sabe que o número de itens vendidos em uma semana, segundo o preço em reais, foi o seguinte:

| Preço | Número de itens vendidos |
|-------|--------------------------|
| 10 | 4 |
| 11 | 4 |
| 13 | 10 |
| 14 | 2 |

E o comerciante deseja saber, qual foi o preço médio da semana??

O calculo é feito da seguinte maneira:

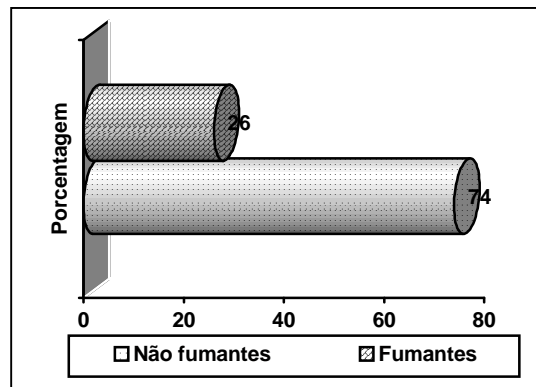
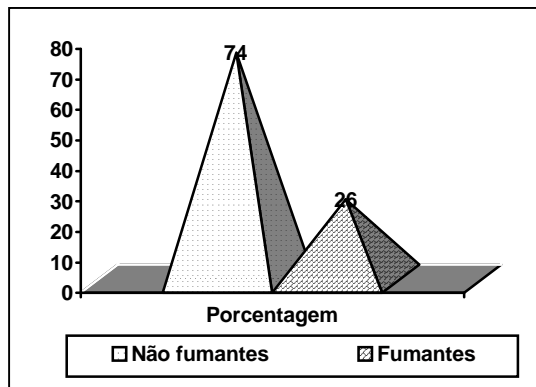
1- Some o número de itens vendidos na semana, para obter o total.

| Preço | Número de itens vendidos |
|-------|--------------------------|
| 10 | 4 |
| 11 | 4 |
| 13 | 10 |
| 14 | 2 |
| Total | 20 |

2- Multiplique preço pelo número de itens vendidos nesse preço.

| Preço | Número de itens vendidos | Produto |
|-------|--------------------------|---------|
| 10 | 4 | 40 |
| 11 | 4 | 44 |
| 13 | 10 | 130 |
| 14 | 2 | 28 |
| Total | 20 | |

3- Some os produtos



Atualmente, com a ajuda dos computadores, os gráficos são ferramentas muito úteis e extremamente precisas, e de fácil execução.

MEDIDAS DE POSIÇÃO

Muitas vezes é necessário um resumo da informação. Veja, o seguinte exemplo:

Um dia Juninho chega da escola, no final do semestre, e sua mãe pergunta:

- *Passou de ano, meu filho??*

Juninho responde:

- *Mãe, tirei 10, 2, 9, 6, 8.*

E a mãe, reclama:

- *Ora filho, passou ou não passou??*

Juninho:

- *Claro que passei, minha média foi !!!!*

Em situações como estas, que faz-se necessário o resumo da informação

Média aritmética – ou simplesmente média: É a soma de todos os valores, dividida pelo número desses valores. Indica-se por \bar{X} (leia xbarra).

Exemplo:

10,2,9,6,8

$$\text{a média é } \frac{10 + 2 + 9 + 6 + 8}{5} = \frac{35}{5} = 7$$

2. Escolha o gráfico de acordo com o tipo de variável

Se a variável for qualitativa, faça:

- ⇒ Gráfico de barras;
- ⇒ Gráfico de setores;

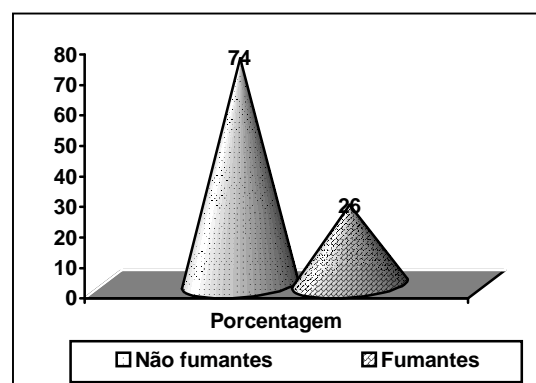
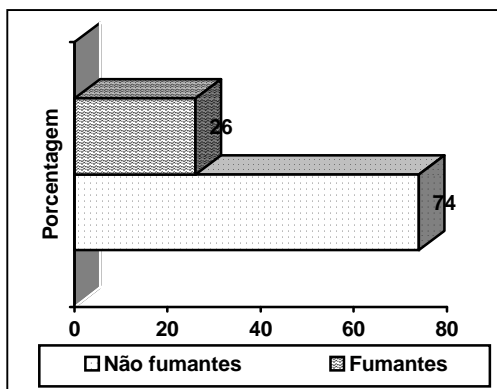
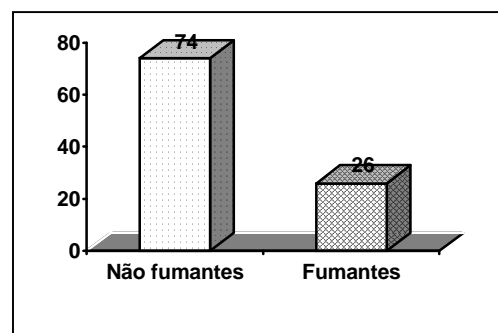
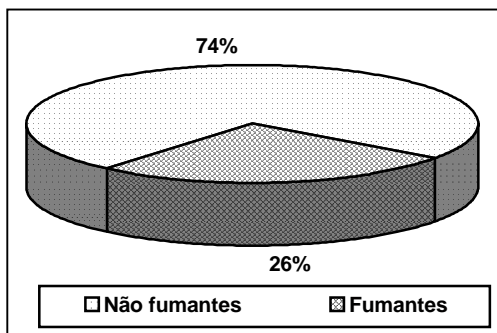
Se a variável for quantitativa, faça:

- ⇒ Histograma
- ⇒ Polígono de frequência

Cuidado! Quando você tem muitas categorias, é melhor não optar pelo gráfico de setores, já que este não fica muito adequado.

Tipos de Gráficos

| Hábito | Frequência | Porcentagem |
|--------------|------------|-------------|
| Não fumantes | 3938 | 74% |
| Fumantes | 1384 | 26% |
| Total | 5322 | 100% |



| | | |
|---------------|------|------|
| Envenenamento | 3010 | 4020 |
| Enforcamento | 2521 | 822 |
| Arma de fogo | 9932 | 1200 |
| Outros | 1125 | 932 |

Calcule:

- o número total de suicídios cometidos por homens e mulheres.
- O método de suicídio mais utilizado pelos homens e o método de suicídio mais utilizado pelas mulheres
- Transforme os dados da tabela em porcentagem.

7. Num grupo de 125 machos e 80 fêmeas, qual é a razão macho/fêmea?

8. Um banco selecionou ao acaso 25 contas de pessoas físicas em uma agência, em determinado dia, obtendo os seguintes saldos em dólares:

| | | | | | | | | |
|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| 89,00 | 701,15 | 112,77 | 159,43 | 848,20 | 948,67 | 245,03 | 190,59 | 607,27 |
| 571,86 | 559,00 | 407,54 | 682,73 | 152,06 | 669,29 | 455,80 | 489,46 | 444,37 |
| 396,32 | 491,00 | 720,23 | 507,63 | 505,18 | 862,22 | 328,68 | 420,77 | 116,47 |
| 842,44 | 134,12 | 246,21 | 520,82 | 474,91 | 212,92 | 710,87 | 755,61 | 772,35 |

Agrupe, por frequência, os dados.

COMO FAZER UM GRÁFICO

1. Examine a variável

As variáveis podem se qualitativas: expressando categorias, características, ou quantitativas, exprimindo quantidade de algo.

As variáveis qualitativas se distribuem em categorias. Você então conta os elementos que caem em cada categoria. Exemplos: Sexo – masculino ou feminino; Credo – católico, protestante, espírita, etc..

As variáveis quantitativas são medidas. Você observa uma medida de cada elemento. Como por exemplo: peso, altura, etc.

- Usando os dados do quadro abaixo, onde aparecem telespectadores e não-telespectadores, reclassificados quanto ao seu grau de realização profissional, calcular: (a) a porcentagem dos que não são telespectadores, porém grandes realizadores; (b) a porcentagem dos telespectadores que são grandes realizadores; (c) a proporção de não telespectadores que são grandes realizadores; (d) a proporção de telespectadores que são grandes realizadores.

| | Situação | |
|------------------|-------------------|---------------|
| | Não-telespectador | Telespectador |
| Realização | | |
| Alta realização | 93 | 46 |
| Baixa realização | 90 | 127 |
| Total | 1783 | 173 |

- Num grupo de 4 sujeitos de alta realização e 24 de baixa realização, qual é a razão dos primeiros para os segundos?
- Faça uma tabela para mostrar que, numa prova de matemática, 2 alunos obtiveram notas 3, 1 aluno obteve 4, 3 alunos obtiveram 5, 4 obtiveram 6, 7 obtiveram 7, e tirar 9 e 1 tirou 10.
- Suponha que se confere grau A aos alunos que obtêm notas no intervalo 9 a 10, grau B aos alunos que obtêm notas no intervalo de 7 a 9, Grau C aos alunos que obtêm notas no intervalo de 5 a 7 e grau D, aos alunos que obtêm notas no intervalo de 0 a 5. Refaça a tabela anterior, dando apenas graus. Escreva o percentual de alunos que recebeu cada grau.
- Faça uma tabela para mostrar que de um total de 852 homens entrevistados sobre determinado assunto, 59 não tinham opinião, 425 eram favoráveis e os demais eram contrários. Das 725 mulheres entrevistadas, 99 tenham opinião, 522 eram favoráveis e as demais contrárias.
- É dada a tabela

| Método | Sexo | |
|--------|-----------|----------|
| | Masculino | Feminino |

| | |
|-------------|--|
| 30 _____ 40 | |
| 40 _____ 50 | |
| 50 _____ 60 | |

2- Organize a tabela

| Classe | Frequência |
|-------------|------------|
| 10 _____ 20 | 3 |
| 20 _____ 30 | 7 |
| 30 _____ 40 | 10 |
| 40 _____ 50 | 6 |
| 50 _____ 60 | 4 |

Resumo:

As variáveis contínuas podem assumir valor num intervalo contínuo. Os dados referentes a tais variáveis dizem dados contínuos;

As variáveis discretas assumem valores inteiros. Os dados discretos são o resultado de contagem do número de itens.

Os dados nominais surgem quando se definem categorias e se conta o número de observações pertencentes a cada categoria.

Os dados por postos consistem de valores relativos atribuídos para denotar ordem: primeiro, segundo, terceiro, quarto, etc.

Foram introduzidas algumas das técnicas básicas usadas pelo pesquisador na organização do emaranhado de dados originais (brutos) que ele coleta de sem entrevistados. Distribuições de frequência para dados que se repetem, e para seqüência de dados que não apresentam valores repetidos.

Para exercitar

| Idade dos empregados de uma Firma | | | | | |
|-----------------------------------|----|----|----|----|----|
| 15 | 30 | 39 | 18 | 33 | 21 |
| 42 | 43 | 49 | 46 | 38 | 29 |
| 59 | 57 | 58 | 35 | 53 | 29 |
| 34 | 39 | 45 | 49 | 43 | 33 |
| 22 | 22 | 35 | 27 | 32 | 19 |

Para organizar os dados em faixas, siga os seguintes passos:

1. Estabeleça os intervalos de grupamento dos dados. Para isso basta calcular, o maior valor – o menor valor amplitude).
2. Decidir qual o número de classes (k) que vai utilizar. É aconselhável tomar de 5 a 15 classes. Menos de 5 classes pode ocultar detalhes importantes dos dados. Mais de 15 classes torna apresentação demasiado detalhada. Calcule \sqrt{n} , sendo n o número de dados total de você tem.
3. Dividir o intervalo encontrado no passo 1, por k , para obter a amplitude de classe (A),
4. Estabeleça os intervalos preliminares, começando com um inteiro logo (LI) abaixo do menor valor dos dados.

A segunda classe é definida por: $(LI)+(A)$

A terceira classe é definida por : $(LI)+(A)+(A)....$

Veja a notação! Nesta classe estão incluídos os indivíduos que têm 10 anos ou mais, contudo, estão excluídos os que têm exatamente 20.

- 1- Conte quantos indivíduos caem em cada classe

| Classe | Frequência |
|----------|------------|
| 10 —— 20 | |
| 20 —— 30 | |

| |
|----|
| 5 |
| 6 |
| 7 |
| 8 |
| 9 |
| 10 |

3- Conte quantas vezes cada número aparece na tabela

| | |
|----|--|
| 5 | |
| 6 | |
| 7 | |
| 8 | |
| 9 | |
| 10 | |

4- Ordene a tabela

| Notas | Frequência |
|-------|------------|
| 5 | 2 |
| 6 | 4 |
| 7 | 6 |
| 8 | 6 |
| 9 | 5 |
| 10 | 1 |

Contudo, os dados também podem ser organizados por faixas. Isso pode ser feito, ou com dados discretos ou contínuos. Entretanto, esta representação pode ser adotada sempre que os dados não apresentam muitas repetições. Como por exemplo em pesquisas, geralmente, a idade é organizada por faixa etária. Vejamos um exemplo, em que conseguimos organizar um quadro com muitos dados, para um quadro mais resumido. veja:

TABELAS DE DISTRIBUIÇÃO DE FREQUÊNCIAS

Para dados discretos, que possuem valores repetidos, espalham-se, às vezes, ao longo de uma extensa amplitude (valor maior – o valor menor), o que torna a distribuição das frequências muito longa, mas, também, difícil de ler. Para solucionar este problemas, tem-se as tabelas de distribuição de frequência. Ou ainda, quando os valores são inteiros e se repetem pode ser organizada uma tabela mais simples denominada tabela de distribuição de frequência.

Exemplo: Sou professora e resolvi avaliar, com base na estatística, as notas dos meus alunos da 6ª série em História, apresentadas no quadro 1.

Quadro 1 – Notas dos Alunos

| | | | | | |
|---|---|---|----|---|---|
| 7 | 8 | 5 | 8 | 6 | 8 |
| 8 | 7 | 6 | 10 | 6 | 9 |
| 9 | 9 | 7 | 7 | 6 | 7 |
| 9 | 9 | 7 | 7 | 6 | 7 |
| 5 | 8 | 8 | 9 | 7 | 9 |

Para construir uma tabela de distribuição de frequências, siga os seguintes passos:

- 1- Procure o número maior e o número menor

Quadro 1 – Notas dos Alunos

| | | | | | |
|----------|---|---|-----------|---|---|
| 7 | 8 | 5 | 8 | 6 | 8 |
| 8 | 7 | 6 | 10 | 6 | 9 |
| 9 | 9 | 7 | 7 | 6 | 7 |
| 5 | 8 | 8 | 9 | 7 | 9 |

O **5** é o menor número, e o **10** é o maior número

- 2- Escreva números inteiros consecutivos em coluna, começando pelo menor, terminando pelo maior

Nível intervalar

O nível intervalar orienta-nos relativamente à ordem das categorias, bem como, nos indica a distância exata entre elas. Escalas intervalares implicam em unidades constantes de medida, como por exemplo: cruzeiros ou centavos, graus Celsius ou Fahrenheit, metros ou centímetros, minutos ou segundos.

Desse modo, uma medida intervalar de preconceito contra os neerlandeses – por exemplo, as respostas a uma série de perguntas sobre esses sujeitos avaliadas numa escala de 0 a 100 (o máximo de preconceito seria 100) – poderia fornecer dados como os que constam na Tabela 1.3 sobre os 9 estudantes.

Pela tabela, observa-se que agora podemos ordenar os alunos, e ainda indicar as distâncias que separam uns dos outros. É possível dizer que Patrícia é a menos preconceituosa da classe, uma vez que ela recebeu o menor escore (a menor nota), e ainda que Patrícia é apenas ligeiramente menos preconceituosa que Osvaldo e Felipe, mas muito menos preconceituosa do que Adriana, e assim por diante.

Dependendo do propósito para o qual o estudo foi feito, seria importante determinar tais informações, as quais, entretanto, não estariam disponíveis no nível ordinal de medida

| Atitude com relação aos Neerlandeses | Escore ¹ |
|--------------------------------------|---------------------|
| Adriana | 98 |
| Maria | 96 |
| Benito | 95 |
| José | 94 |
| Cátia | 22 |
| 21 | 20 |
| Felipe | 15 |
| Pedro | 11 |
| Patrícia | 6 |

¹ Escores mais altos indicam maior preconceito

Tabela 1.1 Atitudes de dez estudantes universitários com relação ao preconceito com um determinado grupo social (Neutrialianos)

| Atitude com relação aos Neutrialianos | Frequência |
|---------------------------------------|------------|
| 1 = Preconceituosos | 5 |
| 2 = Não preconceituosos | 5 |
| Total | 10 |

Nível ordinal

Quando o pesquisador vai além desse nível medida e procura **ordenar** seus sujeitos em função do grau que apresentam de determinada característica, ele está trabalhando com o nível **ordinal** de medida. Exemplo: um pesquisador poderia estar interessado em classificar os indivíduos com referência a *status* sócio-econômico e dispô-lo em classe baixa, classe média, e classe alta. Ou, em lugar de categorizar os alunos de uma dada classe em preconceituosos ou não preconceituosos, ele poderia classificá-los de acordo com o grau de preconceito contra os “neutrialianos” à semelhança do que vai indicado na Tabela 1.2.

| Atitude com relação aos Neutrialianos | Frequência |
|---------------------------------------|--------------------------------------|
| Adriana | Mais preconceituosa (primeiro posto) |
| Maria | . |
| Benito | . |
| José | . |
| Cátia | . |
| Geraldo | . |
| Felipe | . |
| Pedro | ligeiramente preconceituosa |
| Patrícia | menos preconceituosa |

Observe que não é possível atribuir valores a cada aluno, apenas uma posição quanto a sua opinião.

ALGUNS TIPOS DE DADOS

Para o pesquisador é necessário classificar qual o tipo de número considerado, e estes números são classificados por nível de mensuração.

Os número podem ser:

1. para **categorizar** ao nível de mensuração denominado nominal,
2. para **atribuir** postos ou **ordem** ao nível de mensuração denominado ordinal e
3. para **avaliar** ao nível de mensuração denominado intervalar

Nível nominal

O nível nominal envolve simplesmente o ato de nomear ou rotular, em outras palavras, consiste em colocar indivíduos (sujeito experimental) em categorias e contar a frequência com que ocorrem. Para ilustrar, poderíamos usar uma medida de nível nominal para indicar se cada entrevistado tem ou não preconceito com relação a determinado grupo social. Poderíamos submeter 10 alunos de determinada classe a um questionário e verificar que 5 podem ser considerados “preconceituosos” (=1) enquanto que 5 “não-preconceituosos” (= 2).

Exemplo de outras medidas de nível nominal: sexo (feminino versus masculino), classe sócio-econômica (alta e baixa), partido político (A e B), caráter social (voltado “para dentro e tradicional), modo adaptação (conformismo, inovação, ritualismo, fuga, rebeldia), orientação no tempo (presente, passado, e futuro), urbanização (urbano, rural e suburbano).

Devem ter em mente, que a escala nominal, não é gradual, ordenada ou escalonável quanto às qualidades tais como melhor ou pior, ou mais alto, ou mais baixo. O propósito da escala nominal é agrupar os sujeitos em categorias separadas para indicar diferença com referência a uma dada qualidade ou característica.

Soma de quadrados

Os quadrados devem ser somados - $\sum x^2 = (x_1)^2 + (x_2)^2 + (x_3)^2 + (x_4)^2 + \dots + (x_n)^2$

$$\sum x^2 = (x_1)^2 + (x_2)^2 + (x_3)^2 + (x_4)^2 = 2^2 + 7^2 + 9^2 + 6^2 = 4 + 49 + 81 + 36 = 170$$

Soma de produtos

Os produtos são somados - $\sum xy = x_1y_1 + x_2y_2 + x_3y_3 + x_4y_4 + \dots + x_ny_n$

Exemplo:

| | | | |
|---|---|---|---|
| x | 1 | 3 | 1 |
| y | 0 | 9 | 1 |

$$\sum xy = 1.0 + 3.9 + 2.1 = 0 + 27 + 2 = 29$$

Para treinar o que foi aprendido:

Exercício 1:

| | | | | |
|---|---|---|---|---|
| x | 1 | 0 | 3 | 2 |
| y | 2 | 4 | 1 | 3 |

Calcule:

- a) $\sum x$ c) $(\sum x)^2$ d) $\sum x^2$
 e) $\sum y$ f) $(\sum y)^2$ h) $\sum y^2$ g) $\sum xy$

Exercícios 2:

| | | | | | |
|---|---|---|---|---|---|
| X | 1 | 3 | 1 | 4 | 5 |
| Y | 2 | 2 | 0 | 2 | 3 |

Verifique se:

- a) $\sum x = 14$ c) $(\sum x)^2 = 196$ d) $\sum x^2 = 52$
 e) $\sum y = 8$ f) $(\sum y)^2 = 81$ h) $\sum y^2 = 21$ g) $\sum xy = 31$

Exercício 3: Desenvolva das uma das expressões

- a) $\sum_{i=1}^5 x_i$ b) $\sum_{i=1}^5 f_i x_i^2$ c) $\sum 3i$, com $i = 4$ a 8

vendida (em litros) por dia, ou gasolina vendida por hora, a elasticidade de uma tira de borracha – todos são dados contínuos.

Contudo, existem determinados dados que só podem assumir certos valores, em geral inteiros, estes são chamados de dados discretos. Os dados discretos surgem de contagem do número de itens com determinada característica. Exemplos de dados discretos são o número de clientes, de alunos numa sala de aula, de defeitos num carro novo, de acidentes numa fábrica, de paradas de um caminhão, etc.

Antes de começarmos a nos aprofundar mais, na Estatística, vamos nos lembrar de algumas notações...

- Para indicar variáveis usamos as últimas letras do alfabeto como: X, W, Y, Z
- Usamos letras minúsculas para indicar valores que observamos: x, w, y, z

Uma aplicação:

Imagine que uma aluna fez 4 provas... As notas obtidas foram:

$$x_1, x_2, x_3, \text{ e } x_4$$

A soma das notas é dada por:

$$x_1 + x_2 + x_3 + x_4$$

Essa soma também pode ser escrita

$$\sum x \text{ que se lê: "Somatório de x"}$$

Note:

- 1- O símbolo Σ , que se lê "somatório", é uma letra grega sigma maiúscula.
- 2- O símbolo Σ indica que os valores da variável devem ser somados
- 3- É fácil usar o símbolo. Veja: Dados $x_1 = 2$, $x_2 = 7$, $x_3 = 9$, $x_4 = 6$, quanto vale $\sum x$?
- 4- Vale:

$$\sum x = x_1 + x_2 + x_3 + x_4 = 2 + 7 + 9 + 6 = 24$$

Quadrado da soma

A soma é elevada ao quadrado - $(\sum x)^2 = (x_1 + x_2 + x_3 + x_4 + \dots + x_n)^2$

$$(\sum x)^2 = (x_1 + x_2 + x_3 + x_4)^2 = (2 + 7 + 9 + 6)^2 = 24^2 = 576$$

INTRODUÇÃO

OS ESTÁGIOS DA PESQUISA

Testar sistematicamente nossas idéias sobre a natureza da realidade muitas vezes requer uma pesquisa cuidadosamente planejada e executada, onde:

1. o problema a ser estudado é reduzido a uma hipótese testável, ou seja, definir o problema e as hipóteses. (por exemplo “ famílias de um genitor produzem mais delinqüência que famílias de dois genitores”);
2. um conjunto de instrumentos adequado é desenvolvido (por exemplo, são elaborados um questionário ou esquema de uma entrevista);
3. Dados são coletados (isto é, o pesquisador pode ir a campo e fazer uma contagem ou um inquérito);
4. os dados são analisados e comparados com as hipóteses iniciais
5. os resultados de análise são interpretados e comunicados ao público, por exemplo, por meio de conferência ou publicação

Pretende-se com esta apostila tratar mais, os aspectos da pesquisa relacionados com a análise de dados, os quais, depois de coletados ou recolhidos pelo pesquisador, são analisados à luz das hipóteses iniciais. É nesse estágio da pesquisa que os dados brutos são tabulados, calculados, contados, resumidos, reclassificados, comparados ou numa palavra, organizados, a fim de que a precisão ou a validade de nossas hipóteses possam ser testadas.

Na maior parte das vezes, a escolha do processo a utilizar na análise ou descrição de dados estatísticos depende do tipo de dados considerados. Devemos aprender a identificar e a utilizar os vários tipos de dados: discretos, contínuos, nominais, ordinais e intervalares (estes dois últimos, também chamados de postos).

As variáveis que podem assumir qualquer valor num intervalo de valores, são chamadas de contínuas. Características tais como, altura, peso, comprimento, espessura, velocidade, temperatura enquadram-se nesta categoria. Assim, a quantidade de café

ESTATÍSTICA BÁSICA

Daniel Francisco Neyra Castañeda
Bacharel, Licenciado e mestre em Estatística

Julho de 2000