Coursera: Introduction to Statistician

The **histogram** allows to use blocks with different widths

Key point: the areas of the blocks are proportional to frequency.
The histogram gives to two kinds of information:

1. density (crowding)-> the height of the bar
2. percentage (relative frequencies) area = height * width

Box plot (**whisker**)

Boxplot conveys less information than a histogram, but it takes up less space and so is well suited to compare several datasets.
Shows five numbers: smallest, 1st quartile, median, 3rd quartile, largest
(inter-quartile range 3rd-1st quartile)

Scatterplot:

Is sued to depict data that comes as pairs

Std = sqrt((xi-x_bar)^2 / n) or sqrt((xi-x_bar)^2 / n-1)

Complementary rule: p (a not occur) = 1 – P (a occur)

Rule for equally likely outcomes P(A) = 1/ n

A and B are **mutually exclusive** if they cannot occur at the same time.

Addition Rule: P (a or b) = P(a) + P(b)

Conditional probability P（B|A） = P(A and B)/ P(A)

Multiplication rule: P(A and B) = P(A)P(B|A)

Bayes' Rule:

P（B|A） = P(A|B)P(B)/ P(A)

= P(A|B)P(B) / [P(A|B)P(B) + P(A|not B)P(not B)]

Bayesian analysis

The empirical rule:

About 2/3 of the data fall within one std of the mean
About 95% fall within 2 std of the mean
About 99% fall within 3 std of the mean

Standardizing data:
To compute the areas under the normal curve, we first standardize the data:
Z = (height – x_bar) / s -> standardized value or z-score

Standardized data: mean=0 std=1

Binominal Distribution

The binomial formula describes the probability of getting a certain number of successes and failures in an experiment.

N!/ (N-k)!k! * p^k (1-p)^(n-k)

Standard error:
SE = sigma / sqrt(n)

Expected value and std for the sum:

E(Sn) = n * miu
SE(Sn) = sigma * sqrt(n)

Expected value and std for percentages:

E(percentage of 1s) = miu * 100%
SE(percentage of 1s) = sigma / sqrt(n)   * 100%

The law of large numbers:

The law of large numbers states that an observed sample average from a large sample will be close to the true population average and that it will get closer the larger the sample.

The Law of Large Numbers refers to averages (percentages), not sums, as their standard error increases with the sample size.

The Central Limit Theorem

The central limit theorem (CLT) states that the distribution of sample means approximates a normal distribution as the sample size gets larger, regardless of the population's distribution.

Miu = p

Sigma = sqrt[np(1-p)]

The correlation coefficient

The correlation coefficient tells us about the direction and strength of linear relationship.

Confidence intervals via Central Limit Theorem

Estimate +- z SE
e.g. 95% CI z=1.96
90% CI. Z=1.65

The **bootstrap principle** states that we can estimate sigma by its sample version s and still get an approximately correct confidence interval

Example:
We pool 1000 likely voters and find that 58% of them approve the way the president handles his job.
SE = sigma/sqrt(n) * 100%, where sigma = sqrt(p(1-p)), p = proportion of all voters who approve

The bootstrap principle replaces sigma by s = standard deviation of the 0/1 labels in the sample = sqrt(0.58 (1-0.58)) = 0.49
So a 95% CI for p is
58% +- 2 * 0.49 / sqrt(1000)   -> [54.9%, 61.1%]

z-statistic:
z = (observed – expected) / SE

large values of |z| are evidence against H0.
The strength of the evidence is measured by the p-value (observed significance level)

The Monte Carlo Method

A Monte Carlo simulation is **a model used to predict the probability of different outcomes when the intervention of random variables is present**.
Monte Carlo simulations help to explain the impact of risk and uncertainty in prediction and forecasting models.

Bootstrap Confidence Interval

If the sampling distribution of theta is approximately normal, then

[theta – z_a/2 SE(theta), theta + z_a/2 SE(theta)]

Is an approximate (1-a) confidence interval for theta

Exercise:

1. ***We want to use the Monte Carlo method to estimate the probability of getting exactly one ace (one spot) in three rolls of die.***

To simulate three rolls of a die, we draw three times a number at random (with replacement) from 1,2,3,4,5,6. If we get the number `1' exactly once, then we label this trial to be a success.

We repeat this B=1000 times. The proportion of successes in these 1000 trials is our Monte Carlo estimate of the probability in question.

2. ***We want to use the Monte Carlo Method to approximate the standard error of our estimate from Question 1.***

We repeat the whole Monte Carlo simulation done in Question 1 many times (e.g. 2000 times).

Each time we get an estimate of the probability in question. We compute the standard deviation of these 2000 estimates.

3. ***We want to use the bootstrap to estimate the bias of theta^, E(theta^) – theta. Where theta is some function of our population of interest.***
   Theta: population, theta^ = sample.

As usual, we only have access to data from a sample of this population.

The bootstrap plug=in principle suggests to estimate the bias
E(theta^) – t(population)
By
E(theta*^) – t (sample)

E(theta*^) can be approximately by Monte Carlo, resulting in the bootstrap estimate of bias

1/B sum(theta*^ - theta^(sample))

4. ***We want to compute a 90% bootstrap percentile interval for the correlation coefficient based on 32 pairs (X_1,Y_1…,(X_{32},Y_{32})(X1,Y1),…,(X32,Y32). Which of the following is a correct description for doing this?***

Resample 32 pairs (that is, don't break any pairs apart) and compute the correlation coefficient r^*r_* of these 32 pairs.
Repeat B=1000 times to get B bootstrap versions

Chi-Squared Test

1. Testing goodness-of-fit

X^2. = sum (observed – expected) ^ 2 / **expected**

Large values of the chi-square statistic X^2 are evidence of against of H0

The p-value is the right tail of the X^2 distribution with df = number of categories – 1

2. Testing homogeneity

   X^2. = sum (observed – expected) ^ 2 / **expected**

   Df = (no. of columns – 1) (no. of rows – 1)

3 . Testing independence

Comparing several means

The <u>an</u>alysis <u>o</u>f <u>va</u>riance (ANOVA) F-test

Treatment sum of squares
SST = sum_j sum_i (y_j_bar – y_bar) ^ 2 has k-1 df

The **treatment mean square**
MST = SST/ k-1
Measures the variability of the treatment means y_j_bar

The **error sum of squares**
SSE = sum_j sum_i (y_ij – y_j_bar)^2
Has N-k df

The error mean square
MSE = SSE/ N-k

F = MST/ MSE
Under the null hypothesis of equal group means this ratio should be about 1.

It follows a F distribution with k-1 and N-k df
Large value of F suggests the variation between the groups is unusually large. We reject H0.