Forgotten Statistics
Douglas Downing and Jeffrey Clark - Barrons.

Introduction to statistics

- Descriptive statistics → summarizing data.
- Inferential statistics → analyse data by studying a randomly-chosen sample from the population.
- Statistical analysis requires a 'control group' separate from the 'test group' for meaningful conclusions.
- If the test administrators themselves don't know what they are administering (drug or placebo), the survey is 'double blind'.
- Confidence interval → test a hypothesis about specific values of the mean.
- Regression Analysis: test for the presence of a relationship between two or more variables.

Descriptive Statistics

- Mean: $\mu = \bar{x} = \sum_{i=1}^{n} x_i$
- Median: middle value; interpolate for even number of values.
- Frequency histogram (bar diagram).
- Mode: The most frequently-occurring value.
- Normal distribution; the frequencies are symmetric, with a single mode, then the mean, median, and mode coincide at the highest point, which is the highest point of the distribution.
- Bimodal distribution: has two modes.
- If the distribution is uniformly symmetric, the mean and the median will coincide in the middle, but neither is a good indicator of a typical value.
- In an asymmetric distribution with a long tail to the right, the mean will be higher than the median, which will be higher than the mode.
- Variance: measures the degree to which the numbers are spread out.
- Population variance:

$$\mathrm{Var}(x) = \sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n} = \bar{x^2} - \bar{x}^2$$

- Standard deviation σ has the advantage that it is measured in the same units as the original data.
- Note: *sample* variance $s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$.

Probability

- $P(A) = \frac{N(A)}{s}$
- $P(!A) = P(\tilde{A}) = P(A_c) = 1 - P(A)$
- $P(A \,\text{or}\, B) = P(A) + P(B) - P(A \,\text{and}\, B)$
- $P(A \,\text{given}\, B) = P(A|B) = \frac{P(A \,\text{and}\, B)}{P(B)}$
- Note:

A and B mutually exclusive:	$P(A\|B) = 0.$
A $\subset$ B:	$P(A\|B) = \frac{P(A)}{P(B)}.$
A and B independent:	$P(A \,\text{and}\, B) = P(A).P(B)$
	$\Rightarrow P(A\|B) = P(A)$

- If k items from onee list can be combined with m items from another list, there are km possibilities.
- If you make n choices with replacement from a group of m objects, there are m^n possibilities.
- n distinct objects can be arranged $n!$ many ways.
- Permutations of n objects taken r at a time $= n_{P_r} = \frac{n!}{(n-r)!}$.
- Combinations of n objects taken r at a time $= n_{C_r} = \frac{n!}{r!(n-r)!}$.

Random Variable Distributions

- Random variable - discrete or continuous.
- Expected value:

$$E(X) = \mu = \sum i.P(i)$$
$$E(X+Y) = E(X) + E(Y)$$
$$E(cX) = cE(X)$$

- Variance:

$$\text{Var}(X) = \sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n} = E(X^2) - [E(X)]^2$$

where $E(X^2) = i^2 \times P(X = i)$

- x, y independent $\Rightarrow Var(X+Y) = Var(X) + Var(Y)$
- $Var(cX) = c^2 Var(X)$
- The law of large numbers:

$$Var(\bar{x}) = \frac{\sigma^2}{n}, \lim_{n \to \inf} \sigma = 0$$

i.e. as n tends to infinity, the variance of the average tends to zero, as the average value will approach the expected value.

- Special discrete random variable distributions include the binomial, multinomial, Posson, and hypergeometric distributions.

- Continuous random variables:
  The area under the curve represents the probability that the continuous random variable has a value between the curve endpoints.
  The curve is the probability density function.

- Normal Distributions:

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

  μ is the mean/peak of the distrigution

  In order to determine the area under a curve, we can use a table generated for $\mu = 0, \sigma = 1$ (standard normal distribution).

  For a variable X having normal distribution, we create a new variable $z = \frac{X-\mu}{\sigma}$, which has standard normal distribution. We can then look up the probability that z is less than any particular value from the table.

  e.g.:

  $$\begin{aligned} &P(3.5 < A' < 3.6) \\ &= P(\frac{3.5-\mu}{\sigma} < \frac{A'-\mu}{\sigma} < \frac{3.6-\mu}{\sigma}) \\ &= P(N_1 < Z < N_2) \\ &= P(Z < N_2) - P(Z < N_1) \end{aligned}$$

- Continuous random variable distributions related to the normal distributions include the chi-square, t, and f distributions.

Polls, Sampling, and Confidence Intervals

- Hypergeometric distribution (discrete, for sampling without replacement)

- Binomial distributions (discrete, sampling with replacement, easier to calculate).

- In general, if the population size is much larger than the sample size, sampling with and without replacement produce almost identical results, so we can treate the exercise as being with replacement, even though it really isn't (refer to the finite population correction factor).

- We can simplify matters further and approximate the binomial distribuion with the normal distribution. Two major differences are that the normal distribution is continuous, and has negative values.

- Again, as the population size is increased, the normal distribution does a better job of approximating the binomial distribution (refer to the central limit theorem).

- Population proportion: $p = \frac{\text{\# items of interest in population}}{\text{\# items in population}} = \frac{X}{n}$.

- Sample proportion: $\hat{p} = \frac{\text{\# items of interest in sample}}{\text{\# items in sample}} = \frac{X}{n}$.

- Population: $\begin{array}{lll} X_{\text{(normal)}} & \mu = np & \sigma^2 = np(1-p) \\ \hat{p} = \frac{X}{n}_{\text{(also normal)}} & \bar{x} = p & \sigma^2 = \frac{p(1-p)}{n} \\ & & \text{(variance decreases as } n \text{ increases).} \end{array}$

- Confidence interval for the unknown population proportion: getting the sample proportion equal to the population proportion is highly unlikely. We could, however, define an *interval* which would have a high probability of containing the value p.

- $P(\hat{p} - c < p < \hat{p} + c) = 0.95$ (a typical value) $\Rightarrow c = 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

Hypothesis Testing: Unknown Population Parameters

- Probability: Known process → predict outcomes.

- Statistical inference: observe outcomes → predict the nature of the unknown process.

- Estimator: a sample statistic that is being used as an indicator of an unknown paramter (e.g. μ, σ); often designated with a ˆ , e.g. $\hat{\mu}$=estimator for μ.

- An estimator $\hat{A}$ is 'unbiased' if $E(\hat{A}) = A$. A good estimator will also have a small variance.

- An unbiased estimator for the variance is:

$$s = \frac{\sum_{i=1}^{n} X_i - \bar{x}}{x-1}$$

- Hypothesis testing:

- H_0 = null hypothesis = hypothesis being tested, e.g. $\mu = \mu_0$.

- H_a = alternative hypothesis = hypothesis that would hold were H_0 wrong.

	H_0 correct	H_0 wrong
accept H_0	right	type II error
reject H_0	type I error	right

- P(type I error) = level of significance; the lower this is, the stronger the effect of accepting H_0.

- Hypothesis testing procedure:

  1. Decide on the significance level to use (typically 5% or 1%).
  2. Decide on a test statistic to use, such that if H_0 is true, the test statistic will come from a known distribution.
  3. Decide on an acceptance zone and a rejection zone (critical region) for the test statistic.

4. Conduct observations and calculate the test statistic value. If the significance level is set low, we can be quite confident about rejecting the hypothesis.

   The p-value of a test determines the minimum significance level for which we can still reject the hypothesis.

- We first convert the test statistic into the standard normal random variable z:

$$z = \frac{\bar{x} - \mu*}{(\frac{\sigma}{\sqrt{n}})} = \frac{\sqrt{n}(\bar{x} - \mu*)}{\sigma}$$

- Since we don't know σ, we approximate it using s:

$$T = \frac{\bar{x} - \mu*}{(\frac{s}{\sqrt{n}})} = \frac{\sqrt{n}(\bar{x} - \mu*)}{s}$$

- The t-distribution is used for small sample sizes ($n > 60$).

- To use a t-table, we need to use the degrees of freedom = $(n - 1)$. For $n > 60$, we can approximate and use the standard normal distribution.

- Confidence interval for a population mean:

$$\begin{aligned} & P(\bar{x} - c < \mu < \bar{x} + c) = 0.95 \\ & P(-c < \bar{x} - \mu < c) = 0.95 \\ & P(\tfrac{-c\sqrt{n}}{s} < \tfrac{\bar{x}-\mu\sqrt{n}}{s} < \tfrac{c\sqrt{n}}{s}) = 0.95 \\ \text{Let } a = \tfrac{-c\sqrt{n}}{s},\ T = \tfrac{\bar{x}-\mu\sqrt{n}}{s} \quad & \\ & \Rightarrow P(-a < T < a) = 0.95 \end{aligned}$$

- Using degrees of freedom $(n-1)$ to find a from the t-table, we get a=2.145, so $c = \frac{2.145s}{\sqrt{n}}$

- General statement: Hypothesis testing and confidence intervals for means.
  - Assume that the population has a normal distribution or assume that n is large enough that $\bar{x}$ can be assumed to have a normal distribution by the CLT.
  - Assume that the number of items in the population M is much greater than the number of items in the sample population n or that the sample is selected with replacement.
  - Calculate: $\bar{x} = \frac{\sum_{i=1}^{n} X_i}{n}$, $s = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{x})^2}{n-1}}$.
  - Define the random variable $T = \frac{\sqrt{n}(\bar{x}-\mu)}{s}$.
  - This will have a t-distribution with $(n - 1)$ degrees of freedom, if the hypothesis is true.
  - Calculate the end points of the interval: $\bar{x} \pm \frac{a_s}{\sqrt{n}}$, where a is obtained from the t-distribution table using the significance level, and the degrees of freedom.
  - As n increases, the t-distribution better approximates the standard normal distribution, so we can simply use this for $n > 60$.

The relationship question: Regression Analysis

- Regression Analysis: fitting lines that represent patterns of dots.
- We try to find the line that minimizes the sum of several errors.
- The r-squared value indicates how well a line fits:
- r^2 is the percent of variation in y that can be accounted for by variations in x.

  $$r^2 = 1 - \frac{\sum(error)^2 \text{for line}}{\sum(error)^2 \text{for average value}} \qquad (0 \leq r^2 \leq 1)$$

  $r^2 = 1 \rightarrow$ perfect fit.
  $r^2 = 0 \rightarrow$ the variables are completely independent.
- Note: if the slope of the ine is zero, r^2 is undefined, i.e. y is not dependent on x, it is constant.
- Statistical anlysis of regression:
  - e is a random term with mean 0 and an unknown (hopefully small) variance σ.
  - r is the correlation coefficient $(-1 \leq r \leq +1)$.
- Multiple regression is performed when more than one independent variable is involved.

Fallacies and traps

- Beware of graph scales.
- Hypothesis testing is for *unknown* populations using a *randomly* selected sample population.
- For $\sigma = \sqrt{\frac{\sum(x_i-\mu)}{n}}$, $s = \sqrt{\frac{\sum(x_i-\mu)}{(n-1)}}$.
- 'Coincidences' are only amazing if we calculate the probability beforehand. e.g. Calculate the probability that you will see a car with number FLG 927 tomorrow, rather than calculating afterwards that this could have happened.
- We might need to create a dummy variable to obtain matched observations.
- *Statistically significant* does not necessarily imply *significant.*
- Testing many regressions before stating a hypothesis leeaves one open to problems with replicating results.
- $x\%$ of $y = z$ is not the same as $y\%$ of $x = z$
- 95% confidence interval:
  - The mean μ is not a random variable - it has a fixed value that is unknown to us.

- The confidence interval is a random variable.
- There is a 95% chance that the random interval $\bar{x} - \frac{as_z}{\sqrt{n}}$ to $\bar{x} + \frac{as_z}{\sqrt{n}}$ will contain the unknown value of μ.

Introductory Statistics - Jenny Gosling - Pascal Press.

- Estimation: we want to estimate the value of the unknown parameter θ. $\hat{\theta}$ = estimated value of θ
- We usually go for an interval estimate $\hat{\theta}\pm$ Allowable error.
- Confidence coefficient $(1-\alpha)$ the probability that [$\hat{\theta}\pm$ Allowable error] will enclose the true value of θ, e.g. 0.95, 0.99 etc.
- We want:

$$\begin{aligned}
&|\hat{\theta}-\theta| < C\sigma_{\hat{\theta}} \\
&\Rightarrow P[\frac{|\hat{\theta}-\theta|}{\sigma_{\hat{\theta}}} < C] = (1-\alpha) \\
&\Rightarrow P[-C < \frac{\hat{\theta}-\theta}{\sigma_{\hat{\theta}}} < C] = (1-\alpha)
\end{aligned}$$

- If the parent population is normal, then the sampling distribution of $\hat{\theta}$ is approximately normal.
- If $n \geq 30$, then by the Central Limit Theorem, the sampling distribution of $\hat{\theta}$ is approximately normal, regardless of the distribution of the parent population. $\Rightarrow \frac{\hat{\theta}-\theta}{\sigma_{\hat{\theta}}}$ will have a standard normal distribution, and can be denoted by z; we can then use the z-table to find the value of C.
- Example: For $(1-\alpha) = 0.95$, we have:

$$\begin{aligned}
&P[-C < z < C] = 0.95 \\
&\Rightarrow P[-C < z] + P[z < C] = 0.95 \\
&\Rightarrow 2P[z < C] = 0.95 \\
&\Rightarrow P[z < C] = 0.475
\end{aligned}$$

  Looking up 0.475 from the z-table, we get C=1.96, so 95% of all possible $\hat{\theta}$ values lie between $-1.96\sigma_\theta$ and $+1.96\sigma_\theta$ (standard errors) from θ, 2.5% lie below $-1.96\sigma_\theta$, and 2.5% lie above $1.96\sigma_\theta$.

- The standard score (z-score) corresponding to a given raw score in some population represents the relative position of that raw score in the distribution, by measuring the number of standard deviations it lies above or below the mean:
  $z = \frac{x-\mu}{\sigma}$
  This converts the normal distribution curve of x into the standard normal distribution, having $\mu_z = 0, \sigma_z = 1$.
- A $(1-\alpha)100\%$ confidence interval for a parameter, θ assuming:
  - The distribution of $\hat{\theta}$ is approximately normal.

  - The true standard error σ_θ is unknown and is estimated from the sample data by $s_{\hat{\theta}}$.

is of the form:

for $n \geq 30$:

$\hat{\theta} \pm z_{\frac{\alpha}{2}} s_{\hat{\theta}}$

$s_{\hat{\theta}}$ estimates $\sigma_{\hat{\theta}}$
$z_{\frac{\alpha}{2}}$ is the z-value consisting of the top tail area of $\frac{\alpha}{2}$
in the standard normal distribution.

for $n < 30$:

$\hat{\theta} \pm t^{v}_{\frac{\alpha}{2}} s_{\hat{\theta}}$

$s_{\hat{\theta}}$ estimates $\sigma_{\hat{\theta}}$
$t^{v}_{\frac{\alpha}{2}}$ is the t-value consisting of the top tail area of $\frac{\alpha}{2}$
in the t-distribution with degrees of freedom v.

- For estimation of the mean μ:

  Confidence Interval $= \bar{x} \pm z_{\frac{\alpha}{2}\sigma_{\bar{x}}}$, where $\bar{x} = \frac{\sum x}{n}$, $z_{\frac{\alpha}{2}}$ is found from the z-table, and $\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} \approx s_{\bar{x}} = \frac{s_x}{\sqrt{n}}$.

- Confidence interval estimates for specific parameters when comparing two populations:

  - Confidence intervals for the difference betweent the means of two populations:

    - Matched pairs (dependent spamples)

      - $n_x = n_y$
      - the order of the x's and the y's is important to their meaning.

      The easiset way is to look at the differences between the pairs, i.e. reduce a two-sample problem to a single-sample problem. Of course, this procedure is only valid if the data is paired in the first place.

Population 1 (x's)	Population 3 ($d = (x_i - y_i)$	Population 2 (y's)
x	d	y
μ_x	μ_d	μ_y
σ_x	σ_d	σ_y
Sample:		
$\bar{x}$	$\bar{d}$	$\bar{y}$
s_x	s_d	s_y

    - Completely independent samples
      The best point estimate of $(\mu_x - \mu_y)$ is $(\bar{x} - \bar{y})$
      ⇒ The $(1-\alpha)100\%$ confidence interval for the unknown difference $(\mu_x - \mu_y)$ is

$$(\bar{x} - \bar{y}) \pm z_{\frac{\alpha}{2}} \sigma_{(\bar{x}-\bar{y})}$$

$$\sigma_(\bar{x} - \bar{y}) = \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}$$

Chapter 6: Hypothesis testing:

- Every statistical test of hypothesis is made up of four main elements:
  - A null hypothesis H_0; we can reasonably expect that ...
  - An alternative hypothesis; $H_a : \theta \neq \theta_0$.
  - A test statistic; $\frac{\hat{\theta}-\theta}{\sigma_{\hat{\theta}}}$. The smaller this is, the more likely we will be to accept H_0
    ($s_{\hat{\theta}}$ must be used if we don't know $\sigma_{\hat{\theta}}$).
  - A rejection criterion; Reject H_0 if the test statistic exceeds the determined critical value.
- The hypothesis:

One-sided:	$H_0 : \theta \leq \theta_0$.
	$H_a : \theta > \theta_0$.
One-sided:	$H_0 : \theta \geq \theta_0$.
	$H_a : \theta < \theta_0$.
Two-sided:	$H_0 : \theta = \theta_0$.
	$H_a : \theta \neq \theta_0$.

- By decreasing α, we reduce the rejection region. This reduces the chance of making a type I error (reject incorrectly), but increases the chance of making a type II error (accept incorrectly).
  - State the null hypothesis H_0.
  - State the alternative hypothesis H_a.
  - State the significance level of the test, α.
  - Calculate the test statistic using the sample data.
  - State the rejection criterion in terms of the test statistic.
  - Compare the test statistic value with the critical value, and apply the rejection criterion.
  - Write a concluding statement.
- The test statistic is labelled z or t depending on what test we carry out:
  - Sampling distribution of $\hat{\theta}$ is approximately normal, and $\sigma_{\hat{\theta}}$ is known:
    * z-test.
    * $z = \frac{\hat{\theta}-\theta}{\sigma_{\hat{\theta}}}$
  - $n \geq 30$ so sampling distribution of $\hat{\theta}$ is approximately normal (by CLT), and $s_{\hat{\theta}} \approx \sigma_{\hat{\theta}}$ ($\sigma_{\hat{\theta}}$ unknown) holds:
    * z-test.
    * $z = \frac{\hat{\theta}-\theta}{s_{\hat{\theta}}}$
  - $n < 30$: sampling distribution of $\hat{\theta}$ must be approximately normal (since CLT cannot be used):
    * t-test.

        - $t = \frac{\hat{\theta}-\theta}{s_{\hat{\theta}}}$
        - $n < 30$ implies that $s_{\hat{\theta}}$ may not be a good approximation for $\sigma_{\hat{\theta}}$.

- For the z-test:
    - Two-tailed test: $H_0 : \theta \neq \theta_0$. Reject H_0 if $|z| > cvz_{\frac{\alpha}{2}}$.
      For $\alpha = 0.05, cvz_{\frac{\alpha}{2}} = 1.96 \Rightarrow$ we reject H_0 if $|z| > 1.96$.
    - One-tailed test: $H_0 : \theta \leq \theta_0$. Reject H_0 if $z > cvz_\alpha$.
      For $\alpha = 0.05, cvz_\alpha = 1.645 \Rightarrow$ we reject H_0 if $z > 1.645$.
    - One-tailed test: $H_0 : \theta \geq \theta_0$. Reject H_0 if $z < cvz_\alpha$.
      For $\alpha = 0.05, cvz_\alpha = -1.645 \Rightarrow$ we reject H_0 if $z > 1.645$ (by symmetry).
- Reject H_0 if:
  $$|z| > \begin{cases} cvz_{\frac{\alpha}{2}} & \text{two} - \text{sided} \\ cvz_\alpha & \text{one} - \text{sided} \end{cases}$$
- For μ, we use $z = \frac{\bar{x}-\mu_0}{\sigma_x} = \frac{\bar{x}-\mu_0}{s_{\bar{x}}} = \frac{\bar{x}-\mu_0}{(\frac{s_x}{\sqrt{n}})}$
- When comparing two populations:
    - Matched pairs: difference between the means:
        - $H_0 : \mu_d = 0$
          $$H_a : \begin{cases} \mu_d \neq 0 & \text{two} - \text{sided} \\ \mu_d < 0 & \text{one} - \text{sided} \\ \mu_d > 0 & \text{one} - \text{sided} \end{cases}$$
        - If $(n_x = n_y) < 30$, we need to use the t-test:
          $$t = \frac{\bar{d} - \mu_d}{s_{\bar{d}}} = \frac{\bar{d} - 0}{(\frac{s_d}{\sqrt{n}})} = \frac{\bar{d}\sqrt{n}}{s_d}$$
          Reject H_0 if
          $$|t| > \begin{cases} cvt^v_{\frac{\alpha}{2}} & \text{two} - \text{sided} \\ cvt^v_\alpha & \text{one} - \text{sided} \end{cases} \quad \text{where} \quad v = (n-1)$$
    - Unmatched pairs:
        - $\sigma_{(x-y)} = \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}$
        - The sampling distribution of the sample estimate $(x - y)$ is at least approximately normal if:
            - Both parent populations are normal
              or
            - n_x and n_y are both at least 30.
        - $z = \frac{(\bar{x}-\bar{y})-0}{\sigma_{(\bar{x}-\bar{y})}} = \frac{\bar{x}-\bar{y}}{\sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}}}$
          (Use s if σ is unknown).
        - Reject H_0 if:
          $$|z| > \begin{cases} cvz_{\frac{\alpha}{2}} & \text{two} - \text{sided} \\ cvz_\alpha & \text{one} - \text{sided} \end{cases}$$

- For μ, we use $z = \frac{\bar{x}-\mu_0}{\sigma_x} = \frac{\bar{x}-\mu_0}{s_{\bar{x}}} = \frac{\bar{x}-\mu_0}{\frac{s_x}{\sqrt{n}}}$
- Note: Critical z-values may also be found from the last row of the t-table.
- A two-tailed test at significance level α, corresponds to a confidence interval with confidence coefficient $1-\alpha$.
- A one-tailed test at significance level α, corresponds to a confidence interval with confidence coefficient $1-2\alpha$.