Overview & Cheatsheets

Focuses of Class
- Probability: Study of uncertainty
- Statistics: Science of learning from data.

Applications of Class
- Design of Experiments: Understand relationship b/w experiment design & results.
- Quality & Process Control: Identify failures & reduction in quality early.
- Reliability Design: How consistently good is this thing?
- Probabalistic Design

Types of Statistics
- Descriptive: Look back to understand current/old data.
- Inferential: Look forward to predict future (data).
  - Estimation: Make unknowns known w/ some confidance.
    - Point: Provide discrete values w/ a certainty.
    - Interval: Provide continuous distribution some certainty/probability of being true population value. Often chopped up into intervals by certainty.
  - Hypothesis Testing: Determine how credible some statement about a population.

General Terms
- Population / Universe: Items of interest.
- Sample: Subset of population.
- Parameter: Some measure of population
- Statistic: Some measure of sample.

Descriptive Statistics

Descriptive statistics is the intro, childish stuff. It's not super useful or difficult but guides us thru inferential.

Terms

Here, we differentiate b/w the following types of data (T):

- Quantitative: T is infinite. Can be described numerically.
  - Continuous: $|T| = \mathfrak{c}$. Uncountable. (e.g. height of person)
  - Discrete: $|T| \leq \aleph_0$. Countable. (e.g. # of coin flips until head).
- Qualitative: T is finite. Cannot be described numerically.
  - Nominal: Cannot be ordered meaningfully.
  - Ordinal: Can be ordered meaningfully.

Sampling Methods:

- Observational Study: Can determine correlation.
  - Just measure, not interfere
  - Often done after the fact
  - e.g. surveys
- Experiment: Can determine correlation.
  - Deliberately cause certain conditions/treatments, normally separated into groups (w/ at least one control)
  - Measurements done during study
  - e.g. drug trials

Why Sample?

- Population: Entire group of interest.
- Sample: Group used to represent population (normally subset of population).
  - Judgement Sampling: Sampling likely to result in significant difference b/w pop. & samp.
  - Scientific Sampling: Sampling unlikely to result in significant difference b/w pop. & samp.

We sample b/c it is often impractical or impossible to study entire population.

Sample vs Population Symbols

- μ: pop. mean
- σ^2: pop. variance
- σ: pop. standard deviation
- p: pop. proportion (w/ some characteristic)
- N: pop. size
- $\bar{x}$: samp. mean
- s^2: samp. variance
- s: samp. standard deviation
- p_s: samp. proportion
- n: samp. size

Data Presentation/Visualization

Qualitative Data

W/ qualitative data, we're stuck w/ summary data (count, proportion, mode)

- Frequency Table: Table of category ⇒ count or proportion.
- Bar Chart: Count
- Pie Chart: Proportion

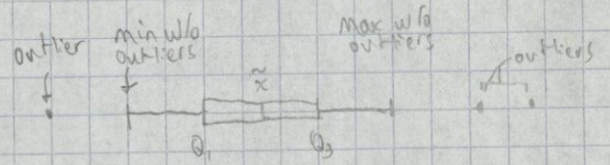
Quantitative Data

Data is diverse, so pick what works best for your scenario.

- Histogram: Bar chart where numbers are evenly binned & there's no space b/w bars.
- Stem & Leaf Plot: You can pick stems & leaves arbitrarily

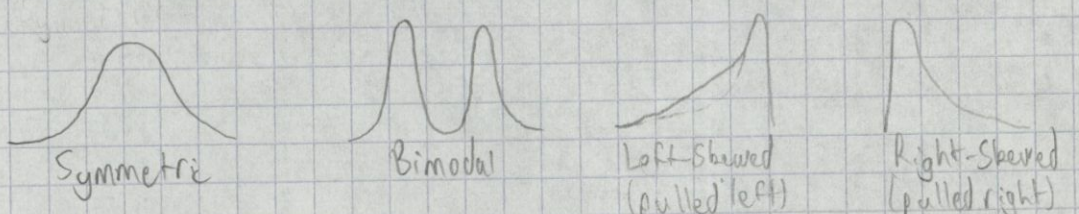
Age in Class	
1	8 8 8 9 9 9 9
2	0 0 0 0 0 1
stems	leaves

- Box & Whisker Plot:



- Frequency Table: Same as above w/ even binning to categorize.

Distribution Shapes



Describing Data/Distributions Numerically

- 5 Number Summary: Min, Q_1, Median ($\tilde{x}$ or Q_2), Q_3, Max

Central Tendency

- Mean ($\bar{x}$): Numerical average. Good for symmetric, numerical data

$$\frac{\sum x_i}{n}$$

- Median ($\tilde{x}$): Middlemost element of data. (Or average if 2 middlemost when n is even). Good for asymmetric numerical data or orderable, non-numeric data.
- Mode: Most frequently occurring datapoint. Good for non-orderable data.

Handling Skewed Data

When we have skewed data set, measuring central tendency becomes more difficult. We can use the following:

- Median ($\tilde{x}$)
- Trim to Mean: We ignore some % at the min & max & calculate the mean w/ the remaining.
  - 10% trim to mean means only consider 10 percentile - 90 percentile for mean.
  - 50% trim to mean is the median!

Variability

Add formulas for population stats! (separate from samples)

Probability

Terms

- Probability: Study of
  - Study of randomness & uncertainty.
  - Measure of randomness.
  - Long term relative frequency of event.
  - $0 \leq \text{prob.} \leq 1$
- Random: State where you can't precisely predict event but it has some trends/rules.
- Experiment: Well-defined process w/ observable outcomes. Here, we deal w/ random ones.
- Sample Point: Single outcome of experiment
- Sample Space (S): Set of all possible sample points / outcomes.
- Event: Subset of sample space.
  - Simple: 1 outcome
  - Compound: 2+ outcomes
    - Think of probability of 1 child being born female from 3 children.

Common Event Descriptions.

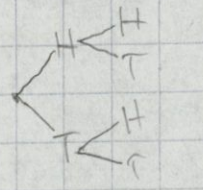
All names E refer to an event on a common sample space S. n(Set) is size of set. P(E) is probability of E in S.

- Complement (E^c): All sample points in S not in E.
- Sample Space of Equally Likely Events:
  $$P(E) = \frac{n(E)}{n(S)}$$
- Mutually Exclusive: For events E_1 & E_2,
  E_1 occurs $\Rightarrow$ E_2 does not occur. } They can both not occur tho!
  E_2 occurs $\Rightarrow$ E_1 does not occur. }
- Independent: For events E_1 & E_2, E_1 occurs gives no info on E_2 & vice versa.

Tree Diagrams

Tree diagrams are a good way of understanding & compound events by breaking it into a tree & travelling around it.

You start at an empty root & branch wherever a random event occurs. These branches have probabilities (but are sometimes left off if all same).

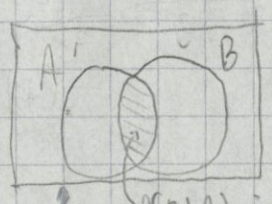


Conditional Probability

Let A & B be any 2 events.
We denote "probability of A given B" as $P(A|B)$.

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

We divide by $P(B)$ to change our scale to account for $P(B)$ having occurred. We take $P(A \cap B)$ to find only when A & B occur.
Basically, we're just reparameterizing $P(A)$ to be in the universe where B occurred.



$P(B|A)$ is this area relative to A circle & not S.

When A & B are independent,

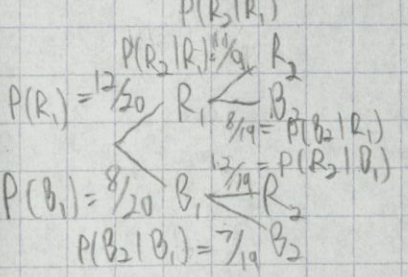
$$P(A|B) = \frac{P(A \cap B)}{P(B)} \overset{A \perp B \text{ independent}}{=} \frac{P(A) \cdot P(B)}{P(B)} = P(A)$$

We get the multiplication rule from $P(A|B)$.

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$
$$\Rightarrow P(A \cap B) = P(A|B)\,P(B)$$

Example

Select two chips sequentially from a pile of 12 red & 8 blue chips, without replacement.



$P(R_1 \cap R_2) = P(R_1)\,P(R_2|R_1)$

= $P(B_2|R_1)$
= $P(R_2|B_1)$

I should've spread this more

* See! The multiplication rule comes from traversing a probability tree.

Independent Events

I've already been talking about this. Oops!

Independent events are events where one occuring says nothing about the other. Mathematically this means:

A & B are independent $\Leftrightarrow$ $P(A|B) = P(A)$ $\Leftrightarrow$ $P(B|A) = P(B)$ $\Leftrightarrow$ $P(A \cap B) = P(A) \cdot P(B)$

[mathematically equivalent]

From multiplication rule & the earlier

For example, your second child being male is independent of your first child being male. However, pulling a red marble out of a finite bag of red & blue marbles is dependent on whether you have drawn a red marble previously.

The following independence & complements are fairly intuitive:

$\Leftrightarrow$ A & B are independent
$\Leftrightarrow$ A & B^c are independent
$\Leftrightarrow$ A^c & B are independent
$\Leftrightarrow$ A^c & B^c are independent

Random Distributions/Variables

You can think of a random variable being a set of tuple(value, probability).

Random variables are variables that have a non-empty set of possible outcomes where each value in the set has corresponding likelihood (probability). We denote this w/ upper case variable names normally.

Let X be the outcome of a coin flip.

$X = \{ H: 0.5, T: 0.5 \}$ ← I'm not sure of the notation, so I came up w/ this

We can have functions act on these random variables of course.

This is important b/c it allows us to deal w/ an entire sample, population, or even family of possible samples analytically at once, rather than dealing w/ individual sample points. We will deal w/ these rather than sample points in the future.

There are 2 types of random variables: discrete or continuous.

* Discrete Random Variables: Finite or countably infinite.
  - X is discrete $\Leftrightarrow \forall x \in \mathbb{N}$
* Continuous Random Variables: Uncountably infinite infinite.
  - X is continuous $\Leftrightarrow \forall x \in \mathbb{R} \wedge x \notin \mathbb{N}$

Discrete Probability Distribution

A finite set of sample points x_i w/ probability prob(x) that satisfy the following

$\forall x \in S \quad 0 < prob(x) \leq 1$
$\forall x \in S \quad \sum prob(x) = 1$

That is, a discrete probability distribution is an exhaustive list of probabilities. This is called the probability mass function.

Measures of Random Distributions/Variables

* Expected Value: Mean of sample space, weighted by probability.
  - $\mu = E(X) = \sum_{x_i \in X} (x_i \cdot prob(x_i))$

* Variance: Variance of sample space, weighted by probability.
  - $\sigma^2 = \frac{\sum_{x_i \in X} ((x_i - \mu)^2 \cdot p(x_i))}{n-1}$

Binomial Distribution

A binomial distribution is a specific, fairly common, well-studied class of situations with the following properties where the number of trials until "success" is the variable of interest:

- Sequence of n repeated trials
- Each trial has 2 possible outcomes (success & failure) ← This is called a "Bernoulli Trial"
- Trials have constant, consistent probabilities
- Trials are independent ← Kinda repetitive.

Essentially, a binomial distribution arises from a sequence of Bernoulli Trials.

Example: Number of coin flips until a head.

The probability of getting k successes in n trials where prob of success is p is given by the probability mass function:

$$\text{prob}(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$$

- p^k: probability of getting k successes
- $(1-p)^{n-k}$: & $n-k$ failures
- $\binom{n}{k}$: number of possibilities for getting k successes from n trials (see "Binomial Coefficient")

* Binomial Coefficient: Number of permutations for ordering k elements of interest from a sequence of n elements.

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

* Note: This is said "n choose k"

Using math & the properties of binomials, we get the following mean & standard deviation

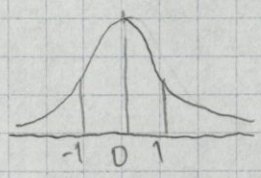
$$\mu = np$$
$$\sigma = \sqrt{np(1-p)} = \sqrt{npq} \text{ where } q = 1-p$$

We won't prove this here b/c "it's not worthwhile."

Continuous Random Variables & Probability Distribution

The function that gives the probability distribution is called the probability density function (PDF). This is different name than that for discrete random variables (called probability mass function (PMF)).

There's a few special ones. Most notably a Gaussian probability distribution (aka a normal distribution).



For all density functions, the area under the curve is 1. That is

$\int_{-\infty}^{\infty} f(x) = 1$ for some probability density function f.

To find probability of sample point being b/w A & B (where A<B), calculate

$P(A \le x \le B) = \int_A^B f(x)\,dx$ for PDF f.

* Note: $P(A \le x \le B) = P(A \le x < B) = P(A < x \le B) = P(A < x < B)$ b/c calculus.

Example:

College professor always finishes his lecture w/in 2 minutes of the bell.

Let X be the time b/w the bell & the end of lecture, given by the PDF

$$f(x) = \begin{cases} kx^2 & : 0 \le x \le 2 \\ 0 & \end{cases}$$

For simplicity we just wrote k. Let's find k since $\int_{-\infty}^{\infty} f = 1$ for PDF f.

$\int_{-\infty}^{\infty} f(x)\,dx = 1$
$\Rightarrow k\int_0^2 x^2\,dx = 1$
$\Rightarrow k\left[\frac{1}{3}x^3\right]_0^2 = 1$
$\Rightarrow k\left(\frac{8}{3} - 0\right) = 1$
$\Rightarrow k = \frac{3}{8}$

$\therefore f(x) = \begin{cases} \frac{3}{8}x^2 & ; 0 \le x \le 2 \\ 0 & \end{cases}$

Find $P(x \le 1)$

$P(x \le 1) = \int_{-\infty}^{1} f(x)\,dx = \frac{3}{8}\int_0^1 x^2\,dx = \frac{3}{8}\left[\frac{1}{3}x^3\right]_0^1 = \frac{1}{8}\left[x^3\right]_0^1 = \frac{1}{8}$

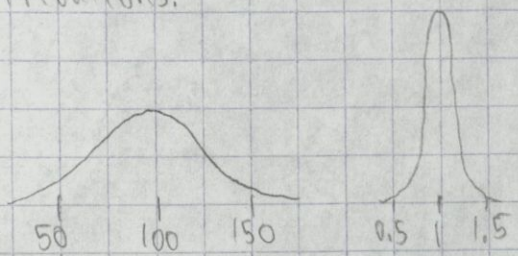
Find $P(0.5 < x < 1.5)$

$P(0.5 < x < 1.5) = \int_{0.5}^{1.5} f(x)\,dx = \frac{3}{8}\int_{0.5}^{1.5} x^2\,dx = \frac{1}{8}\left[x^3\right]_{0.5}^{1.5} = \frac{1}{8}\left(\frac{27}{8} - \frac{1}{8}\right) = \frac{1}{8}\left(\frac{26}{8}\right) = \boxed{\frac{13}{32}}$

Now, in this class we will cover Normal, T, & F continuous distributions.

Normal/Gaussian Distribution

Normal distributions are a family (or set!) of symmetric, centrally focused distributions.



They have similar shapes!

We can completely describe an element from this set w/ 2 numbers: the mean (μ) & standard deviation (σ). We notate these so:

$X \sim N(\mu = \mu_0, \sigma = \sigma_0)$ — Really: $f_X(x) = \frac{1}{\sqrt{2\pi}\cdot\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ Neat!

means X is a normal distribution w/ mean μ_0 & standard deviation σ_0.

* Note: $\mu_0 \in \mathbb{R}$ & $\sigma_0 \in \mathbb{R} > \sigma_0 > 0$

We have an empirical rule or 68-95-99.7 rule, which states:

68% of data is w/in 1 std. dev.
95% of data is w/in 2 std. dev.
99.7% of data is w/in 3 std. dev.

We (semi) arbitrarily choose a Standard Normal Distribution (or "Curve") / Z Curve of $N(\mu=0, \sigma=1)$. This existed initially to simplify calculation before computers (using calculation tables), but now exists for historical reasons.

It is still useful for intuitively comparing elements from different normal distributions. We use something called z-score for this.

The z-score is how many standard deviations you are displaced from the mean. In other words, it is what happens if you normalize your distribution/datapoint w/ the z curve. We calculate it so:

$$z\text{-score}(x) = \frac{x-\mu}{\sigma}$$

Showing $E(z)=0$ & $V(z)=1$. I'm not convinced by this.

$$E(z) = E\left(\frac{x-\mu}{\sigma}\right) = \frac{E(x)-\mu}{\sigma} = \frac{\mu-\mu}{\sigma} = 0$$

($E(z)$: mean (μ))

$$V(z) = V\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sigma^2} V(x-\mu) = \frac{1}{\sigma^2}(V(x)-0) = \frac{\sigma^2}{\sigma^2} = 1$$

↑ variance (σ^2)

R Code

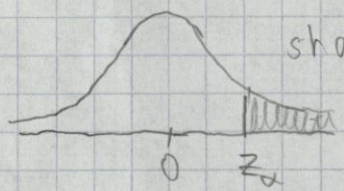
We'll be using the following commonly for continuous random distributions

- rDIST(n, ...): Generate n random points from the given (distribution) DIST using parameters ... for the distribution.
- pDIST(x, ...): Find P(DIST(...) ≤ x). Inverse of pDIST. ← think probability/proportion
- qDIST(p, ...): Find x ∋ P(DIST(...) ≤ x) = p. Inverse of qDIST.

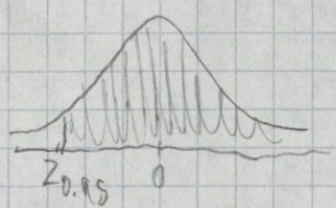
For normal distributions, DIST = "norm" & ... = "mean = ..., sd = ...".

We also use z_α notation. This is simply read as z-score such that α is to the right of the Z curve. In other words

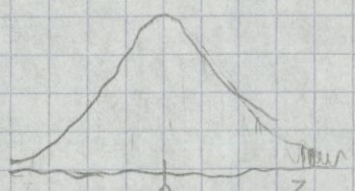
$$z_\alpha = z \ni P(Z \geq z_\alpha) = \alpha$$



Draw $z_{0.95}$



Draw $z_{0.01}$



To Find z_α in R:

z_α = qnorm(1-α)

Inferential Statistics

Inferential statistics is the most applicable part of statistics. It focuses on:

- Hypothesis Testing: Determine whether statement/hypothesis about population is reasonable.
- Estimation: Try to describe/get some info about a population from a sample.
  - Point Estimation: Give single best guess.
  - Interval Estimation: Give probability distribution to possibilities. Often highlight some specific interval.
    - Generally the possibilities' probability distribution is approximately normal.

Sampling Distributions

A sampling distribution is a meta distribution, which shows the variability (& distribution) of some statistic of possible samples of a given size. You form it by considering all possible samples (weighted by likelihood) your sampling method produces. (You can also consider a sample of samples.) These are used a lot in inferential statistics to understand how representative your sample is of the population.

Here, we will mostly consider sample mean ($\bar{x}$) b/c it is easiest to understand. We also mostly consider simple random samples (SRS) b/c there we assume equal likelihood.

*Note: Sampling distributions are smoother than individual samples b/c for samples of size n, the sampling distribution is N

Z Test/Distribution

Where n is sample size & $\bar{x}_i$ is $\bar{x}$ for sample i.

$$\mu_{\bar{x}} = \frac{\sum_{i=1}^{n} \bar{x}_i}{n} = \mu \qquad \sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$$

$$\sigma_{\bar{x}} = \sqrt{\frac{\sum_{i=1}^{n} (\bar{x}_i - \mu_{\bar{x}})^2}{n}} = \frac{\sigma}{\sqrt{n}} \leftarrow \text{This is also called standard error (SE)}$$

*Note: Sampling distributions are smoother than individual samples b/c, for samples of size n, the sampling distribution's size is

N^n for sampling w/ replacement

$\binom{N}{n} = \frac{N!}{n!(N-n)!}$ for sampling w/o replacement.

Because sample mean ($\bar{x}$) is really a random variable often written $\bar{X}$, where $\bar{X}$ depends on population size (N) & sample size (n). (Also depends on sampling method but we assume SRS here.)

(Some books use $N(\mu, \sigma^2)$. We use $N(\mu, \sigma)$ here.)

Central Limit Theorem (CLT)

this is what allows us to use the z-test

The central limit theorem here is a big boy for inferential statistics that allows us to do everything.

Central Limit Theorem: If you take a large random sample ($n \geq 30$), from any population, the sample mean is approximately normal w/ a mean of $\mu_{\bar{x}} = \mu$ & standard deviation of $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.

(Make sure the CLT applies (e.g. SRS w/ $n>30$) if you're not given a normally distributed population.)

$$\bar{X} \sim N(\mu_{\bar{x}}, \sigma_{\bar{x}}) = N(\mu, \frac{\sigma}{\sqrt{n}})$$

The central limit theorem need not apply for normal distributions. This is b/c the sampling distribution is already exactly normal.

The CLT is great b/c it allows us to deal w/ any distribution, not just normal distributions. It basically forces a normal (sampling) distribution out of a non-normal distribution. Remember the bimodal distribution & taking random samples of different sizes from AP statistics?

Estimation w/ Confidence Intervals

Now we're just getting into the good stuff. Note we use the z-test as a representative example.

Trivially, $\bar{x}$ is a point estimate for μ & s for σ. But how confident are we in this? To quantify this, we will use our meta-distribution the sampling distribution to understand how (un)likely our results were.

Here, we'll do this w/ a confidence interval. We construct this by assuming our sample mean & sample variance are true for the population. We then construct a sampling distribution for our sample size (n). This is called the confidence interval.

(we could also use real values)

We normally want a confidence interval at a certain confidence level $(1-\alpha)$. To find this, take the lower & upper bounds for the center $1-\alpha$ proportion from the confidence interval. These are called the critical values or bounds of our confidence interval.

What does α mean? It means that when we run this experiment, α proportion of the time our confidence interval will miss the true mean. $1-\alpha$ proportion of the time we'll capture it.

We won't prove this here b/c it's "not worth it." But, intuitively, it should make sense that you can go from the true mean & standard deviation to find how likely our results were. This process of going from true mean & std. dev. to our results, semi-intuitively, doesn't care about what direction you go. Therefore, our process at least isn't insane.

We have terms other than confidence level $(1-\alpha)$ to describe our confidence interval. We have margin of error which is the magnitude b/w one critical value & the mean.

Summary

To construct a confidence interval at a $1-\alpha$ confidence level, our critical values are

critical value

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

Alternatively, using R for lack of a better syntax

critical value

Low CV = qnorm($\frac{\alpha}{2}$, $\bar{x}$, $\frac{s}{\sqrt{n}}$)
High CV = qnorm($1-\frac{\alpha}{2}$, $\bar{x}$, $\frac{s}{\sqrt{n}}$)

Confidence Intervals in General

For normally distributed sampling distributions for some statistic

Bounds = point estimate $\pm$ CV · SE = point estimate $\pm$ ME
CV (Critical value) or proper multiplier derived uniquely
SE (Standard Error) derived uniquely
ME (Margin of Error) = CV · SE

Unqualified standard error is just standard deviation of statistic

Interpreting Confidence Interval

It is important to interpret a confidence interval. (And, in general, you should interpret the results of any inferential statistics).

Here we say something like

> We are x% confident that the population mean, μ lies between
> Low CV & High CV. This means that, in repeated sampling of the same
> population, x% of all intervals constructed the same way will contain μ.

Student's t Test & Distribution

The t distribution is an alternative to using the CLT & normal distributions. The t distribution depends on the degrees of freedom (df). It is overall better than the CLT, especially at low sample sizes, but is more conservative.

$df = n-1$

The t test is great for unknown pop. standard deviations (σ)

Comparison w/ Normal Distribution

The t distribution is parameterized by df (degrees of freedom).

The t distribution is wider.

The t distribution approaches the normal distribution as $n \to \infty$.

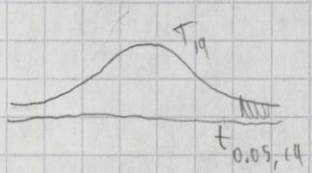
Confidence Intervals w/ t Distribution

This is the same as for using the CLT & normal distributions, except t instead of z.

$$CV = \bar{x} \pm t_{\alpha/2, df}\left(\frac{s}{\sqrt{n}}\right)$$

Our α-notation is the same but we also include df.

$$\therefore \quad t_{\alpha, df} = t \Rightarrow P(T_{df} > t) = \alpha$$



R Code & t Distributions

Recall $z_\alpha = qnorm(1-\alpha)$

Shockingly $t_\alpha = qt(1-\alpha, df)$ *Note: $df = n-1$

Examples

$$t_{0.025,24} = qt(0.975, 24)$$

$$t_{0.1,4} = qt(0.9, 4)$$

Interpreting t Tests

t tests are interpreted exactly like normal distribution tests w/ CLT. This works b/c we were generic when we said "interval constructed the same way".

Preconditions for t Tests

The t-test has pretty liberal requirements

1) For $n<15$, the sample must be symmetric, have a single peak, & have no outliers.
2) For $n \geq 15$, the sample doesn't have outliers or strong skew.

Sampling Distributions w/ Proportion

Previously, we have only dealt w/ means. Now let's deal w/ proportions!

When dealing w/ proportions (e.g. what proportion of people are left-handed), we model it as a binomial distribution where the Bernoulli trials are people are in your proportion or not.

Here, we assume the population as much larger than our sample (so p is constant) & thus model our true proportion as p (chance of success) & n as the sample size. This should be fairly intuitive.

↓ you do other things for non-large pop

Recall, for binomial distribution X
- p is probability of success,
- n is number of trials
- $E(X) = np$ ← expected value
- $V(X) = np(1-p)$ ← variance
- $sd(X) = \sqrt{np(1-p)}$

Symbols/Notation
- p is true/pop. proportion
- $\hat{p}$ is sample proportion.
- X is number of successes in sample
- n is sample size

Deriving $\mu_{\hat{p}}$ & $\sigma_{\hat{p}}$

Recall the rules on expected value & variance:

$$E(a) = a$$
$$E(aX) = aE(X)$$
$$V(a) = 0$$
$$V(aX) = a^2 X$$

We define sample proportion ($\hat{p}$) thus:

$$\hat{p} = \frac{X}{n}$$

Now, let's derive our boys $\mu_{\hat{p}}$ & $\sigma_{\hat{p}}$.

$$\mu_{\hat{p}} = E(\hat{p}) = E\left(\frac{X}{n}\right) = \frac{1}{n} E(X) = \frac{1}{n} \cdot np = p$$
$$\therefore \boxed{\mu_{\hat{p}} = p}$$

$$\sigma^2_{\hat{p}} = V(\hat{p}) = V\left(\frac{X}{n}\right) = \frac{1}{n^2} V(X) = \frac{1}{n^2} \cdot np(1-p) = \frac{p(1-p)}{n}$$
$$\therefore \boxed{\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}}$$

*Note: This is approximately normally distributed similar to how proportions themselves are.

Applying This to Confidence Intervals

Annoying Abbreviations
CV = critical value
SE = standard error
ME = margin of error.

Recall the general form for confidence interval's is

Bounds = point estimate $\pm$ CV · SE (or ME = CV · SE)

Since we're using a normal approximation, we use a z-distribution/test.

$\therefore$ Bounds $= \hat{p} \pm z_{\alpha/2}\,\sigma_{\hat{p}} = \hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Preconditions

We can only do a z-test about proportion if

- The expected successes & failures are large* ($n \cdot \hat{p} > 5$ & $n(1-\hat{p}) > 5$) — *so your probabilities don't significantly change
- The sample is a random sample
- Individual observations are independent

If we don't meet the preconditions, we can often

Hypothesis Testing ← * We have these restrictions on what we can say b/c we assume H_0

Oftentimes, people make statements & we want to test them.

In hypothesis testing, we have a statement & we want to determine if it is unreasonable.

A hypothesis is a statement/belief about the population. In testing, we have a null hypothesis (H_0) which we hold to be true. It is what we want to prove. We also have our alternate hypothesis (H_a), which is the opposite of what we want to prove.

(We want to disprove the null hypothesis.)

We can only ever reject the null hypothesis or fail to reject. Since our H_a is the converse of H_0 & H_0 is what we want to prove, disproving H_0 thus indicates H_a is correct & we are wrong. → to reject H_0 as unreasonable, either our H_0 or H_a may be correct. If we fail

(H_0 must contain =.
H_a cannot contain =.)

Examples

If we have some sample w/ $\bar{x} = 20$, $s = 2$, $n = 100$. We want to test the statement that $\mu = 50$ w/ a 5% level of significance.

$H_0: \mu = 50$
$H_a: \mu \neq 50$

We assume $\mu = 50$ & $\sigma = 2$. (from sample)

$\therefore \bar{x} \sim N(\mu_{\bar{x}} = \mu = 50,\ \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2}{\sqrt{100}} = 0.2)$

We now find the probability of getting $\bar{x} = 20$.

If this probability is $< 5\%$, we reject H_0. Otherwise, we fail to reject.

Types of Errors

Here, we have 2 types of errors.

- Type I Error: Reject H_0 when H_0 is true.
- Type II Error: Fail to reject H_0 when H_0 is false.

In certain scenarios, one of the types of errors is more common. We can tweak our test to be right/wrong the ideal amount of time.

	Actual Situation	
Decision	H_0 is True	H_0 is False
Fail to Reject H_0	$(1-\alpha)$	Type II Error (β)
Reject H_0	Type I Error (α)	Power $(1-\beta)$

In this class, we arbitrarily choose to fix α & then use that fixed α to find β

Level of Significance: α
Power: $1-\beta$

Critical Value Approach

We can prove things in 2 ways. Previously, I talked about the p-value approach. We'll get more into that later.

What "direction" H_0 goes changes how we interpret our level of significance. The direction refers to the rejection direction.

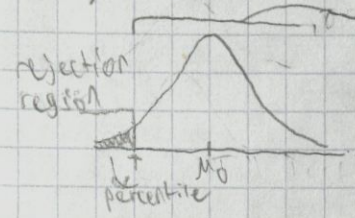
- Lower Tail Test: We consider too low failure.
  $H_0: \mu \geq \mu_0$ Area of Lower Rejection $= \alpha$
  $H_a: \mu < \mu_0$
- Upper Tail Test: We consider too high failure.
  $H_0: \mu \leq \mu_0$ Area of Upper Rejection $= \alpha$
  $H_a: \mu > \mu_0$
- Two Tail Test: We consider too low & too high failure.
  $H_0: \mu = \mu_0$ Area of Lower Rejection $= \alpha/2$
  $H_a: \mu \neq \mu_0$ Area of Upper Rejection $= \alpha/2$

reinterpret relative to p-value approach.

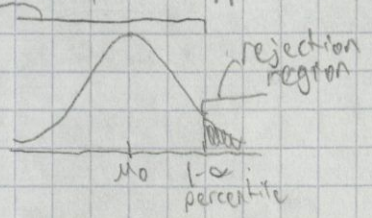
What's all this "rejection region" nonsense? Well, we can reinterpret hypothesis testing by constructing a range of sample statistics which back up / don't reject your H_0. The range not in this acceptable

range, we call it the rejection region.

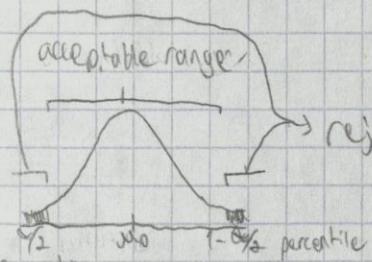
Here, I'll construct $\bar{X}$ if H_0 is true, for lower, upper, & two tail cases.



Lower Tail Test



Upper Tail Test



Two Tail Test

* We always assume the worst case for μ, where $\mu = \mu_0$

For $P(\text{Test I}) = \alpha$ (aka level of significance), assume $\alpha = 0.05 = 5\%$ b/c that is standard.

As you can see, we are more conservative in two tail tests. This is b/c we "are saying more" / "know less" if we're doing a two tail test. It's also just convention.

Procedure

Here, we'll follow a pretty strict procedure for doing tests on μ.

Our null hypothesis (H_0) is always either $\mu \geq \mu_0$, $\mu \leq \mu_0$, or $\mu = \mu_0$. & our alternate hypothesis (H_a) is the converse.

Recall

$\frac{\bar{X} - \mu}{s/\sqrt{n}} \sim N(0,1)$ for $n \geq 30$ (CLT)

$\frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$ for $n \leq 30$ when pop. is not normal

We assume H_0 is true (i.e. $\mu = \mu_0$). Then, we split based on type of test. In general, we find z-score of our observation, assuming H_0, & see whether it is in the acceptable range or not.

Lower Tail Test: Reject H_0 if $\frac{\bar{x} - \mu_0}{s/\sqrt{n}} \leq -z_\alpha$

Upper Tail Test: Reject H_0 if $\frac{\bar{x} - \mu_0}{s/\sqrt{n}} \geq z_\alpha$

Two Tail Test: Reject H_0 if $\frac{\bar{x} - \mu_0}{s/\sqrt{n}} \leq -z_{\alpha/2}$ or $\frac{\bar{x} - \mu_0}{s/\sqrt{n}} \geq z_{\alpha/2}$

Note: You replace z w/ t_{n-1} if you have $n < 30$ & pop. is not normal.

Generalizing Procedure

Simplifying our earlier work, using lower tail test WLOG,

$$\frac{\bar{x} - \mu_0}{s/\sqrt{n}} \leq z_\alpha$$

$$\bar{x} - \mu_0 \leq z_\alpha \frac{s}{\sqrt{n}}$$

$$\bar{x} \leq \mu_0 + z_\alpha \frac{s}{\sqrt{n}}$$

$$\bar{x} \leq \mu_0 + (CV)(SE)$$

$$\bar{x} \leq \mu_0 + ME$$

Essentially, we are finding $z_{obs} \geq z_\alpha$ is true

You can easily derive similar results for the other tests.

Notice that this now matches our intuitionist argument using areas of acceptance & rejection. We won't normally use this form b/c it is more confusing. However, it is beneficial to see.

Constructing Hypotheses

If we want to prove something, we put it as the H_a. This is b/c, if we make H_0 what we want to prove, we'll never be able to do anything but reject it.

By making H_a what we want to show & then constructing H_0 to be the converse of H_a, if we reject H_0 our H_a gains more credibility. This allows us to side-step not being able to prove things & only disprove things. We aren't proving but we are giving H_a credibility.

P-Value Approach

As opposed to the critical value approach, we can find the probability of getting sample assuming H_0. If the probability (p-value) is below a certain level of significance (α), we say H_0 is probably wrong.

This is completely equivalent to the critical value approach, but gives you different intuition behind the process.

Essentially, we are finding

- For upper tail test $P(Z > z_{obs})$
- For lower tail test $P(Z < z_{obs})$
- For 2-tail test $2 \cdot P(Z > |z_{obs}|)$ ← b/c we want probability we're farther away (either higher or lower)

Tests w/ Proportions

So far, we've only done hypothesis testing w/ population mean. Now, we'll extend our method to means.

First, recall

- p is our population proportion.
- $\hat{p}$ is our sample proportion.
- $E(\hat{p}) = p$

Likewise $q = 1 - p$

- $\text{var}(\hat{p}) = \frac{\hat{p} \cdot \hat{q}}{n}$
- $sd(\hat{p}) = \sqrt{\frac{\hat{p} \cdot \hat{q}}{n}}$

Restrictions

The restrictions on this test come from us modelling $\hat{p}$ using a normal distribution. As such, we have the same restrictions as we did earlier.

- $np \geq 5$ for a large number of successes
- $nq \geq 5$ for a large number of failures.

Since we assume the worst case H_0, $p = p_0$ & $q = q_0$.

This will come up later.

Using the Test

Since we are approximating the sampling proportions as normal distributions, we use a z test.

$$z_{obs} = \frac{\hat{p} - E(\hat{p})}{sd(\hat{p})} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 \cdot q_0}{n}}} \sim N(0,1)$$

Recall that we are using $p = p_0$ & $q = q_0$ from assuming the worst case for H_0

Then, we do the same breakdown, critical values, & p-values as earlier

Test Type	Hypotheses	Reject When	P-Value
Upper Tail Test	$H_0: p \leq p_0$ $H_a: p > p_0$	$z_{obs} > z_\alpha$	$P(Z > z_{obs})$
Lower Tail Test	$H_0: p \geq p_0$ $H_a: p < p_0$	$z_{obs} < -z_\alpha$	$P(Z < z_{obs})$
Two Tail Test	$H_0: p = p_0$ $H_a: p \neq p_0$	$\lvert z_{obs} \rvert > z_{\alpha/2}$	$2 \cdot P(Z > \lvert z_{obs} \rvert)$

Hypothesis Test for Two Populations

Here, we have two population 1 & 2 where σ

- σ_1 unknown
- σ_2 unknown
- $\sigma_1 \neq \sigma_2$
- $n_1 \geq 30$
- $n_2 \geq 30$

Basically, we'll be doing a z test where the hypotheses look like

$H_0: \mu_1 - \mu_2 \leq a$ OR $H_0: \mu_1 - \mu_2 \geq a$ OR $H_0: \mu_1 - \mu_2 = a$
$H_a: \mu_1 - \mu_2 > a$ $H_a: \mu_1 - \mu_2 < a$ $H_a: \mu_1 - \mu_2 \neq a$

To do this, we'll be using the arithmetic of random variables & "way to quantify the relation of 2 random variables.

Covariance

Covariance is a way to quantify the relationship of 2 random variables. Height & weight have relatively high covariance & movie preference & food preference are fairly low.

We won't go into calculating this here, but we should now

$cov(X, Y) = 0 \iff X$ & Y are independent.

$cov(X, Y) > 0 \iff X$ & Y have a positive relationship.

$cov(X, Y) < 0 \iff X$ & Y have a negative relationship.

Small Covariance $\iff$ Little Relationship

Large Covariance $\iff$ Significant Relationship

We can use this to define correlation which is like covariance but unitless, so you can compare them.

$$corr(x, y) = \frac{cov(x, y)}{sd(x) \cdot sd(y)}$$

Algebra of Random Variables

$$E(X_1 \pm X_2) = E(X_1) \pm E(X_2)$$
$$var(X_1 \pm X_2) = var(X_1) + var(X_2) \pm 2\,cov(X_1, X_2)$$

Applying this to sampling distributions, we can make some simplifications. Specifically, sample means will always be independent. (Imagine the relationship essentially "cancelling out".) This gives us

$$E(\bar{X}_1 \pm \bar{X}_2) = E(\bar{X}_1) \pm E(\bar{X}_2)$$
$$var(\bar{X}_1 \pm \bar{X}_2) = var(\bar{X}_1) + var(\bar{X}_2)$$

We can rewrite this into more familiar terms of σ & μ

$$\mu_{\bar{x}+\bar{y}} = \mu_{\bar{x}} + \mu_{\bar{y}} = \mu_x + \mu_y$$

$$\mu_{\bar{x}-\bar{y}} = \mu_{\bar{x}} - \mu_{\bar{y}} = \mu_x - \mu_y$$

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

This gives us z values of

$$z_{\bar{x}+\bar{y}} = \frac{(\bar{x}+\bar{y}) - \mu_{\bar{x}+\bar{y}}}{\sigma_{\bar{x}+\bar{y}}} = \frac{(\bar{x}+\bar{y}) - (\mu_x + \mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}$$ ← These are our z_{obs} values!

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

Hypothesis Tests & Confidence Intervals

Now that we have a way to get the z-value of our combined distribution of 2 populations, we can trivially use our earlier critical value & p-value test w/ the same procedure, once we have our z-value. Same w/ confidence intervals.

Applications

Neat! Now, why would we want to do any of this?

Combining two populations is most important when we want to compare two populations.

For example, we want to find if the mean of population x is greater than population y. While we could frame it as

$H_0: \mu_x \leq \mu_y$
$H_0: \mu_x > \mu_y$

You'll find it more obvious on how exactly you compare them framing it as

$H_0: \mu_x - \mu_y \leq 0$
$H_a: \mu_x - \mu_y > 0$

Also, sometimes you just want to combine two populations & analyze them. Or, maybe break down a population.

Size & Tests

How do we determine to use a z-test or a t-test? Normally you do z-test when $n \geq 30$ & t-test $n \leq 30$.

Here, we use the minimum sample size from one of the populations as n & do the above when $\sigma_1 \neq \sigma_2$.

To simplify things, we split up into 3 cases (showing confidence intervals here). Here, x & y are our two populations.

Case n_x & $n_y \geq 30$ & $\sigma_1 \neq \sigma_2$

$$(\bar{x}-\bar{y}) \pm z_{\alpha/2}\sqrt{\frac{s_x^2}{n_x}+\frac{s_y^2}{n_y}}$$

In real life, a z-test is rarely used b/c it is strictly worse than a t-test. A t-test approaches a z-test as n becomes large.

Case n_x or $n_y < 30$ & $\sigma_1 \neq \sigma_2$

$$(\bar{x}-\bar{y}) \pm t_{\alpha/2,df}\sqrt{\frac{s_x^2}{n}+\frac{s_y^2}{n}}$$

IS THIS RIGHT?

where $df = \min(n_x-1, n_y-1)$

Case n_x or $n_y < 30$ & $\sigma_1 = \sigma_2$

$$(\bar{x}-\bar{y}) \pm t_{\alpha/2,df}\sqrt{s_p^2\left(\frac{1}{n_x}+\frac{1}{n_y}\right)}$$

where $df = m+n-2$ ← sum of their individual degrees of freedom

$$s_p^2 = \frac{(n_x-1)s_x^2+(n_y-1)s_y^2}{n_x+n_y-2}$$ ← this is called the pooled variance

Hypothesis Testing w/ Paired Data (aka. Changes)

When test data comes in pairs, it doesn't make sense to treat the data as 2 separate populations b/c they're absolutely not independent. Additionally, since the data comes in pairs, we can make simplifications

TO simplify, we combine the data by using the difference b/w pairs & using this as your data. We then treat it as a single population.

We often subscript the variables w/ d for difference.

Multiple Proportions

Recall:
For a binomial random variable B w/ n trials & p prob. of success
- $E(B) = np$
- $V(B) = np(1-p)$

Recall combined variance formulas

$V(X+Y) = V(X) + V(Y) + 2\text{cov}(X,Y)$ ← should make sense that dependent values tend to compound.

$V(X-Y) = V(X) + V(Y) - 2\text{cov}(X,Y)$ ← should make sense that dependent values tend to not be very different.

Recall variance scalars

$V(aX) = a^2 V(X)$

$V(a) = 0$

* Hats ($\hat{}$) on top of symbols notates an estimate. (We don't follow this for means b/c of historical reasons.)

We can use all these rules to define rules for combining $\hat{p}_1$ & $\hat{p}_2$, where $\hat{p}$ is the sample proportion.

$V(\hat{p}_1 - \hat{p}_2) = V\left(\frac{x_1}{n_1}\right) + V\left(\frac{x_2}{n_2}\right)$ where 1 & 2 are independent

x is the sample success count

n is the sample size

$= \frac{V(x_1)}{n_1^2} + \frac{V(x_2)}{n_2^2}$

$= \frac{n_1 p_1 q_1}{n_1^2} + \frac{n_2 p_2 q_2}{n_2^2}$

$= \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}$

Recall: General Confidence Interval

(test statistic) ± (margin of error) (critical value)

$\therefore \boxed{sd(\hat{p}_1 - \hat{p}_2) = SE = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}}$ & $p = \frac{x}{n}$, $q = \frac{n-x}{n}$ $\Rightarrow \boxed{SE = \sqrt{\frac{x_1(n_1-x_1)}{n_1^3} + \frac{x_2(n_2-x_2)}{n_2^3}}}$

We can use this standard error (SE) to normalize our $\hat{p}_1$ & $\hat{p}_2$ doing

$$\frac{\Delta\hat{p} - \Delta p}{SE} = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}}$$

Now that we've tackled the variance & standard error of multiple proportions, we've gotten most of it out of the way.

We still need to handle when $\sigma_1 = \sigma_2$, since we can't do variance separately. We have to pool the standard deviation. In this case,

$$sd(\hat{p}_1 - \hat{p}_2) = SE = \sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

where $\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$ ← Freshman sum / combined proportion (basically average w/ weights)

$\hat{q} = 1 - \hat{p}$

Hypothesis Testing w/ Arbitrarily Many Groups

In this class, we will only cover how to test the following simple hypotheses, since it gets very complicated.

Well, proportions are a fairly simple extension.

H_0: $\mu_1 = \mu_2 = \dots = \mu_n$
H_a: Not H_0 (at least one is different)

To do this, we do an ANalysis Of VAriance (ANOVA).

One-Way ANOVA

Abstractly, if the variation b/w groups is greater than the variation w/in groups, the means are different. This should make sense.

We won't go into detail, but we can make a summary table

where a is the number of groups
n is the number of people in a group

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square (aka Variance)	F_{obs}
b/w groups → Treatment	$a-1$	SS_{Tr}	$MS_{Tr} = SS_{Tr}/(a-1)$	$\frac{MS_{Tr}}{MS_E}$
w/in groups → Error	$a(n-1)$	SS_E	$MS_E = SS_E/(a(n-1))$	
Total	$an-1$	$SS_{Tr}+SS_E$		

* Note: We act as though we always have the same number in a group

This should look very familiar, where s^2 is $\frac{\sum(\bar{x}-\mu)}{df}$.

SS_E is sometimes simply called Error.

In general, the closer the test statistic is to 1, the more equal the groups. The farther from 1, the less equal.

Our goal for all this nonsense is to get our test statistic (f_{obs}). We then compare the expected f-distribution & our f_{obs}. An f-distribution is parameterized by the treatment's degrees of freedom (df_T) & the error's degrees of freedom (df_E).

f_{df_T, df_E}

In R, we can find the area to the left of some value x in an F-distribution parameterized by df_T & df_E using

$pf(x, df_T, df_E)$

To do a hypothesis test, in this class we'll always use the p-value approach using the above code & the test statistic f_{obs} we showed how to get earlier.

Just to reiterate, the p-value is

$$p\text{-value} = P(f_{df_T, df_E} > f_{obs})$$

It is totally possible to do the critical value approach! In fact, you could probably do it right now. But, for this class, we'll always use the p-value approach.

* Note: MS_E (the mean squared error) estimates the population variance of each population. This means we assume all populations have the same variance.

This parallels our pooled variance (s_p^2) from earlier!

When the number in each group is unknown, we find the total number of individuals across groups to find $df_T + df_E$. We then find df_T as normal & solve for df_E.

Suppose we have 3 groups. 2 w/ 21 people. 1 w/ 32.

We know $df_T + df_E = 74 - 1 = 73$

$df_T = 3 - 1 = 2$

$\therefore df_E = 73 - 2 = 71$

Here we talk about deterministic & probabilistic models. You already know this so I skipped it. Essentially, probabilistic has some random error. (ε normally)

Simple / Single Linear Regression

Simple linear regression tries to relate 2 variables via a line. Concretely, where x & y are variables, find β_0 & β_1 such that (where ε is error)

$$y = \beta_0 + \beta_1 x + \varepsilon$$

We do this by finding a line that minimizes the sum of squares / variation. We do this by doing calculus. Here's the equation formally.

$$\hat{\beta}_1 = \frac{\sum(x_i y_i) - \frac{(\sum x_i)(\sum y_i)}{n}}{\sum(x_i^2) - \frac{(\sum x_i)^2}{n}}$$

The hat means these are just estimates!

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ ← just run it back!

* Note: This only works if we have (x,y) pairs (duh!)

We have an alternative form of β_1 w/ the correlation coefficient (r) that measures how strongly correlated the two variables are.

$$\beta_1 = r \cdot \frac{s_y}{s_x} \quad \text{where} \quad -1 \le r \le 1$$

We often want to measure how good a prediction is/was. We can do this w/ residual error. The residual error is just the difference b/w the predicted value ($\hat{y}_i$) & actual value (y_i)

$$\text{residual} = y_i - \hat{y}_i$$

The hat means it's an estimate (like normal)

How do you get the predicted values ($\hat{y}_i$) by doing

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 i$$

we're using the predicted β_0 & β_1 & ignoring error!

Review w/ Pete ☺ doesn't really fit here but I'm low on paper

Type I Error - Rejecting Truth
Type II Error - Accepting False

Test for Means (1-Sample)

In all cases σ is unknown. (If σ is known & pop. is norm, a z-test is perfect.)

Normal Population? — Yes → t-test

No → n? — > 30 → z-test ← Central Limit Theorem

≤ 30 → you're screwed

you'll be told whether you can assume normality

*Normality trumps sample size

Test for Proportions (1-Sample)

We can test proportions the same way we do means, since sample proportion distributions look like sample mean distributions given

$$\begin{cases} n > 30 \\ n p_0 \geq 10 \\ n(1-p_0) > 10 \end{cases}$$ ← heuristics

All we do is define standard deviation (s) as

$$s = \sqrt{\frac{p_0(1-p_0)}{n}}$$ ← p_0 rather than $\hat{p}$ b/c it has nicer properties

Tests for Means (2-Sample)

In all cases σ is unknown. If both are known, the a z-test is perfect.

Population Normal? — Yes → $\sigma_1 \stackrel{?}{=} \sigma_2$ — = → t-test equal variance

$\sigma_1 \stackrel{?}{=} \sigma_2$ — ≠ → z-test equal variance

Population Normal? — No → n? — ≥30 → z-test

n? — ≤30 → you're boned

Recall $s^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2$

Simple Linear Regression

We assume x & y are related by the following, where β_0 & β_1 are constants & ε is a constant / not-varying normal random variable

$$y = \beta_0 + \beta_1 x + \varepsilon$$

How do we minimize the error? The statistics community has broadly settled on minimizing the square residual.

We can estimate these w/

$$\beta_1 = \frac{\Sigma(x_i y_i) - \frac{(\Sigma x_i)(\Sigma y_i)}{n}}{\Sigma(x_i^2) - \frac{[\Sigma x_i]^2}{n}}$$ ← these will be given

r^2 is "the total variance in y explained by x." ← if you sum all r^2, you can get larger than 1.

Confidence Intervals

Confidence intervals have the same flow chart for hypothesis tests, since they're pretty similar. We normally do dual bounds confidence intervals, but you can do only one bound.

Experimental Design

Terms

- Response Variable: Variable which is affected by some other variable.
  - aka: Dependent Variable.
- Explanatory Variable: Variable which affects some other variable.
  - aka: Independent Variable.
- Observational Study: Simply measure something w/o influence. Track a subjects variables of interest to find <u>correlation</u>.
  - Passive
- Experiment: Assign some treatment to subjects & measure your treatment's influence (normally before & after). Meant to find <u>causation</u>.
  - Treatment: Some change in default behavior / course of action. A collection of factors at certain levels.
  - Control Treatment: No-op change.
  - Factors: List of explanatory variables being changed.
  - Level: Value of a factor.
  - Experimental Unit / Subject: Object/individual upon which the experiment is done.
- Mixed Study: Some factors are assigned. Others are merely observed.
- Experimental Error: Variability in response variable w/in the same treatment group.
  - Having experimental error does <u>not</u> mean your experiment is bad. However, it can hint that your treatments are ineffectual.
- Lurking Variable: Some other variable that is not one of your explanatory variables.
  - Example: Children w/ night lights tend to have worse eye sight. This is b/c the parents who have poor eyesight are more likely to buy night lights. The parents' eyesight is a lurking variable.
- Confounding Variables: When multiple variables tend to occur together, making their effects hard to separate.

Principles of Experimental Design

When designing experiments, we <u>control</u> for lurking variables, <u>randomize</u> to reduce bias, & <u>replicate</u> to produce all possible combinations.