MATH 326

Probability & Statistics II

Tom Carty – Spring 2022

TeX'ed by Matthew Wilder (BU '23)

Contents

Common Abbreviations and Notations

dist'n = Distribution
r.v. = Random variable
DOF = degrees of freedom
d.f. = degrees of freedom
MGF = moment generating function
mgf = moment generating function
pdf = probability distribution function
CLT = Central Limit Theorem
iid = Independent Identically Distributed
MSE = Mean Square Error
MoM = Method of Moments
MVUE = Minimum Variance Unbiased Estimator
MLE = Maximum Likelihood Estimator
μ = mean (average)
σ = standard deviation
σ^2 = variance
$\bar{X}$ = Sample mean
S = Sample standard deviation
S^2 = Sample variance
$N(\mu, \sigma^2)$ = Normal distribution with: mean = μ, variance = σ^2
Z = Z-score

Chapter 7

Sampling Distributions and the Central Limit Theorem

Chapter 7 Review

Our first goal in Math 326 is to learn how to estimate global parameters of "population" like μ and σ. To do this, we need to understand how the random variable we use to estimate them are distributed. For example:

$$\text{Parameters} \begin{cases} \mu \\ \sigma^2 \end{cases} \qquad \text{Statistics} \begin{cases} \bar{X} = \dfrac{1}{n}\sum X_i \\ S^2 = \dfrac{1}{n-1}\sum \left(X_i - \bar{X}\right)^2 \end{cases}$$

7.2 Sample Means

① **Sample means:**

Theorem 7.1: Let $Y_1, \ldots, Y_n$ be a random sample from $N(\mu, \sigma^2)$ then $\bar{Y}$ is distributed by $N\left(\mu, \dfrac{\sigma^2}{n}\right)$. The distribution of sample means, $\bar{Y}$, is also Normal.

Reason: Linearity of Expectation and Variance properties

Discussion: Recall in working with $N(\mu, \sigma^2)$, we learned it was easier to standardize everything via Z-scores:

$$Z = \frac{x - \mu}{\sigma}$$

and Z is distributed $N(0, 1)$, the **Standard Normal Distribution.**

② **Sample variance:**

Recall standard deviation σ is a measure of the spread of the random variable and it's derived from $\left(Y_i - \bar{Y}\right)^2$ terms. In all math, we normalize to take the "units" out of things.

$$\underbrace{U_i = \frac{Y_i - \bar{Y}}{S}}_{\text{Data Driven = Stat}} \approx \underbrace{\frac{Y_i - \mu}{\sigma} = Z_i}_{\text{Not a stat}}$$

Reason: Z_i depends on unknown population parameters, nonetheless, it makes it easier to pretend that we start here.

Theorem 7.2 If $Y_1, \ldots, Y_n$ are a random sample of $N(\mu, \sigma^2)$. Then,

$$U = \sum Z_i^2 = \sum \left(\frac{Y_i - \mu}{\sigma} \right)^2$$

has a χ^2 distribution with n degrees of freedom (df).

Recall: χ^2 distribution is a Gamma $\left(\frac{\nu}{2}, 2 \right)$ where $\nu = \text{df}$ (degrees of freedom).

Reason: In old homework (325), ${Z_i}^2$ is χ^2 with $\text{df} = 1$. By product of mgf, $\sum {Z_i}^2$ is χ^2 with $\text{df} = n$. Now to get sample variance, we do some algebra:

$$\begin{aligned} & & S^2 &= \frac{1}{n-1} \sum (Y_i - \bar{Y})^2 \\ & & &\approx \frac{1}{n-1} \sum (Y_i - \mu)^2 \\ \iff & & \frac{S^2}{\sigma^2} &\approx \frac{1}{n-1} \sum \left(\frac{Y_i - \mu}{\sigma} \right)^2 \\ \iff & & \frac{S^2(n-1)}{\sigma^2} &\approx \sum {Z_i}^2 \end{aligned}$$

Showing it is okay to replace $\bar{Y}$ with μ is the point of the proof of the following theorem:

Theorem 7.3 (Fisher's Theorem) The distribution of sample variance S^2

If $Y_1, \ldots, Y_n$ is a random sample from $N(\mu, \sigma^2)$. Then,

① $\dfrac{S^2(n-1)}{\sigma^2}$ has χ^2 distribution with $(n-1)$ degrees of freedom.

② $\bar{Y}$ and S^2 are independent random variables.

③ t-distribution and F-distribution

$$\underbrace{\frac{Y_i - \mu}{\sigma}}_{\text{Normal}} \approx \underbrace{\frac{Y_i - \mu}{S}}_{\text{t-dist'n}} \approx \underbrace{\frac{Y_i - \bar{Y}}{S}}_{\text{Statistic}}$$

Address when to use what is the point of this class

Skip definitions (such as moments) for the time being

Also, The Law of Large Numbers. Used to prove the Central Limit Theorem.

7.4 The Central Limit Theorem

The reason Normal distributions play an outsized role in applied statistics is that the distribution of $\bar{Y}$ can be made *nearly normal*, no matter the underlying distribution of Y (Does not need to start "life" Normal). So *nearly* normal, that we just pretend it is.

Theorem 7.4: The Central Limit Theorem (CLT)

Let $Y_1, \ldots, Y_n$ be independent and identically distributed (iid) random variables with

$$\mathrm{E}(Y_i) = \mu \qquad \text{and} \qquad \mathrm{Var}(Y_i) = \sigma^2 < \infty$$

(Need finite variance for CLT to hold)

Define

$$U_n = \frac{\bar{Y} - \mu}{\sqrt{\frac{\sigma^2}{n}}} = \frac{\sum(Y_i) - n\mu}{\sigma\sqrt{n}}.$$

The CLT says the distribution function of U_n converges to $N(0,1)$ as $n \to \infty$.

Big idea: The distribution $\bar{Y}$ can be thought of as $N\left(\mu, \frac{\sigma^2}{n}\right)$.

Definition: The **support of f** is the domain on which f is non-zero.

Example: Let $\bar{X}$ denote the mean of a random sample of size $n = 15$, from the distribution whose pdf is

$$f(x) = \frac{3}{2}x^2, \qquad x \in [-1, 1]$$

Can be shown that

$$\mu = \mathrm{E}[X] = 0 \qquad \text{and} \qquad \sigma^2 = \mathrm{E}\left[(X-\mu)^2\right] = \frac{3}{5}$$

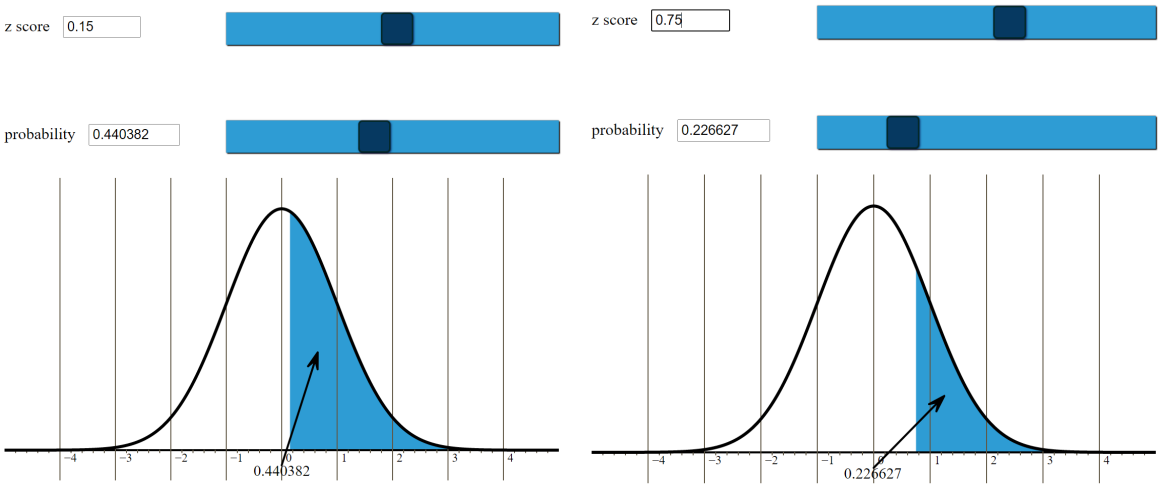
To compute $\Pr\left(0.03 \le \bar{X} \le 0.15\right)$, we use the CLT and assume $\bar{X}$ is distributed by

$$N\left(\mu, \frac{\sigma^2}{n}\right) = N\left(0, \frac{3/5}{15}\right) = N\left(0, \frac{1}{25}\right).$$

Then

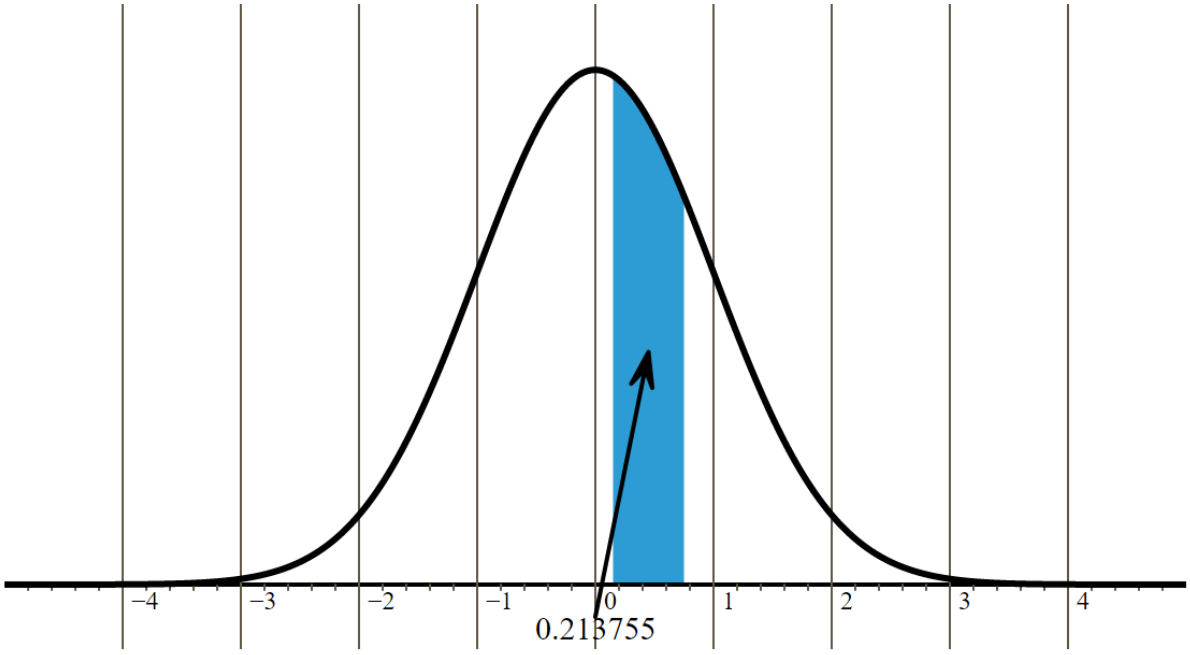
$$\begin{aligned}
\Pr\left(0.03 \le \bar{X} \le 0.15\right) &= \Pr\left(\frac{0.03-\mu}{\sigma} \le \frac{\bar{X}-\mu}{\sigma} \le \frac{0.15-\mu}{\sigma}\right) \\
&= \Pr\left(\frac{0.03-0}{\sqrt{1/25}} \le Z \le \frac{0.15-0}{\sqrt{1/25}}\right) \\
&= \Pr\left(\frac{0.03}{\sqrt{1/25}} \le Z \le \frac{0.15}{\sqrt{1/25}}\right) \\
&= \Pr(0.15 \le Z \le 0.75) \\
&= \text{Table4}(0.15) - \text{Table4}(0.75) \\
&= 0.4404 - 0.2266 \\
&= 0.2138
\end{aligned}$$

Recall: Table 4 gives upper tail probabilities.



Therefore,

$$\begin{aligned}\text{Table4}(0.15) - \text{Table4}(0.75) &= \Pr(Z > 0.15) - \Pr(Z > 0.75) \\ &= 0.4404 - 0.2266 \\ &= 0.2138\end{aligned}$$



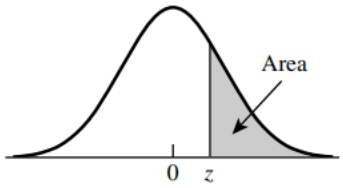
Discussion: About n. Suprisingly, "large" n doesn't need to be that large. Usually for any random variable X and any distribution, then

$$\boxed{n \geq 30 \quad \text{is enough.}}$$

That is, when $n \geq 30$, the CLT says $N\left(\mu, \frac{\sigma^2}{n}\right)$ yields a good approximation of $\bar{X}$.

When X is symmetric, unimodal, and continuous, $n = 4$ or $n = 5$ is often enough.

Table 4 Normal Curve Areas
Standard normal probability in right-hand tail
(for negative values of *z*, areas are found by symmetry)



	Second decimal place of z									
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0722	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0352	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
2.9	.0019	.0018	.0017	.0017	.0016	.0016	.0015	.0015	.0014	.0014
3.0	.00135									
3.5	.000 233									
4.0	.000 031 7									
4.5	.000 003 40									
5.0	.000 000 287									

Chapter 8

Estimation

8.1 An Estimator

We are seeking an unknown population parameter. General name of θ (e.g. μ, σ^2, ρ, ...).

The rule used to approximate or guess θ is called an **estimator**. Any guess based on observations is called an **estimate**.

Example: Want $\theta = \mu$.

Estimator: $\underbrace{\bar{X} = \frac{1}{n}\sum x_i}_{\text{a function}}$

Estimate: Given 7 observations, $Y_1, \ldots, Y_7$, then $\bar{Y} = \frac{1}{7}\sum_1^7 Y_i$ is an estimate.

This is an example of a **point** estimator.

These are also **interval estimators**.

Example: Confidence interval, $\theta = \mu$.

$$\bar{X} - SZ_x \leq \mu \leq \bar{X} + SZ_x \qquad \Longleftrightarrow \qquad |\bar{X} - \mu| \leq SZ_x$$

8.2 The Bias and Mean Square Error of Point Estimators

A start at trying to determine if a point estimator is any good.

Let $\widehat{\theta}$ be a point estimator for a parameter θ. (e.g. $\theta = \mu, \quad \widehat{\theta} = \bar{X}$)

① One goal of a good estimator is that $E(\widehat{\theta}) = \theta$ (e.g. by CLT, $\mathrm{E}\left(\bar{X}\right) = \mu$)
But this is not always the case.

Definition: If $\mathrm{E}[\widehat{\theta}] = \theta$, we say that $\widehat{\theta}$ is an **unbiased** point estimator.
(e.g. $\bar{X}$ is unbiased).

If $\widehat{\theta}$ is a biased point estimator, we define the **bias** to be

$$\mathrm{bias}(\widehat{\theta}) = \mathrm{E}[\widehat{\theta}] - \theta$$

② Another goal of a good estimator might be that its "spread" of observations is tightly packed (talking about variance), and hopefully near θ.

$$\text{spread} \implies \text{variance}$$

Definition: The **mean square error** of $\widehat{\theta}$ is

$$MSE(\widehat{\theta}) = E\left[(\widehat{\theta} - \theta)^2\right]$$

Recall: $V[X] = E[X^2] - (E[X])^2$

Corollary:

$$\mathrm{MSE}(\widehat{\theta}) = \mathrm{Var}(\widehat{\theta}) + \left[\mathrm{bias}(\widehat{\theta})\right]^2$$

Proof.

$$\begin{aligned}
\mathrm{MSE}(\widehat{\theta}) &= \mathrm{E}[\widehat{\theta}^2 - 2\widehat{\theta}\theta + \theta^2] \\
&= \mathrm{E}[\widehat{\theta}^2] - 2\,\mathrm{E}[\widehat{\theta}\theta] + \mathrm{E}[\theta^2] \\
&= \mathrm{E}[\widehat{\theta}^2] - 2\,\mathrm{E}[\widehat{\theta}]\theta + \theta^2 \\
&= \mathrm{E}[\widehat{\theta}^2] - (\mathrm{E}[\widehat{\theta}])^2 + (\mathrm{E}[\widehat{\theta}])^2 - 2\,\mathrm{E}[\widehat{\theta}]\theta + \theta^2 \\
&= \mathrm{Var}(\widehat{\theta}) + (\mathrm{E}[\widehat{\theta}] - \theta)^2 \\
&= \mathrm{Var}(\widehat{\theta}) + (\mathrm{bias}(\widehat{\theta}))^2
\end{aligned}$$

□

Remark: We have decomposed MSE by:

$$\mathrm{MSE}(\widehat{\theta}) = \underbrace{\mathrm{Var}(\widehat{\theta})}_{\text{Precision}} + \underbrace{\mathrm{bias}(\widehat{\theta})^2}_{\text{Accuracy}}$$

Idea of a "best" estimator is tricky. We'd like MSE to be as small as possible, but this is an impossible problem because MSE depends on the unknown $\widehat{\theta}$.

Example: (Population proportions). Want p.

To estimate, let

$$Y = \begin{cases} 1 & \text{"success"} \\ 0 & \text{"failure"} \end{cases}$$

and $\{Y_i\}$ be an iid random sample.

Estimator 1: $\widehat{p} = \dfrac{1}{n}\sum Y_i = \bar{Y}$

This sums the $\frac{\text{people said yes}}{\text{total people}}$

Recall: $\sum Y_i$ is binom(n, p).

And for a Binomial Distribution, we know

$$\mathrm{E}[X] = np \qquad \text{and} \qquad \mathrm{Var}(X) = npq = np(1-p)$$

Bias? Then what is the E$[\widehat{p}]$?

$$\mathrm{E}[\widehat{p}] = \mathrm{E}\left(\frac{1}{n}\sum Y_i\right) = \frac{1}{n}\mathrm{E}\left(\sum Y_i\right) = \frac{1}{n}\cdot np = p$$

Because E$[\widehat{p}] = p$, then $\widehat{p}$ is an unbiased estimator. For MSE$(\widehat{p})$, we need Var$[\widehat{p}]$:

$$\mathrm{Var}[\widehat{p}] = \mathrm{Var}\left(\frac{1}{n}\sum Y_i\right) = \left(\frac{1}{n}\right)^2 \mathrm{Var}\left(\sum Y_i\right) = \left(\frac{1}{n}\right)^2 np(1-p) = \frac{p(1-p)}{n}$$

Therefore,

$$\begin{aligned} \mathrm{MSE}(\widehat{p}) &= \mathrm{Var}(\widehat{p}) - \mathrm{bias}(\widehat{p})^2 \\ &= \frac{p(1-p)}{n} + \underbrace{0^2}_{\text{"unbiased"}} \end{aligned}$$

Estimator 2: $\widetilde{p} = \dfrac{\sum_1^n Y_i + 1}{n+2}$

$$n = 1, \qquad y = \begin{cases} 1 \\ 0 \end{cases} \qquad \widetilde{p} = \begin{cases} 2/3 \\ 1/3 \end{cases}$$

$$n = 2, \qquad y = \begin{cases} 2 \\ 1 \\ 0 \end{cases} \qquad \widetilde{p} = \begin{cases} 3/4 \\ 2/4 \\ 1/4 \end{cases}$$

$$\mathrm{E}[\widetilde{p}] = \frac{\mathrm{E}(Y_i) + \mathrm{E}(1)}{n+2} = \frac{np+1}{n+2} = p + \frac{1-2p}{n+2} \neq p$$

Hence, $\widetilde{p}$ is biased. And,

$$\text{bias}(\widetilde{p}) = \text{E}[\widetilde{p}] - p = p + \frac{1-2p}{n+2} - p = \frac{1-2p}{n+2}$$

Calc question: What happens to $\text{E}[\widetilde{p}]$ as sample size grows? ($\text{E}[\widetilde{p}] \to p$)

Need

$$\begin{aligned}
\text{Var}\,(\widetilde{p}) &= \text{Var}\left(\frac{\sum_1^n Y_i + 1}{n+2}\right) \\
&= \text{Var}\left(\frac{\sum_1^n Y_i}{n+2} + \frac{1}{n+2}\right) \\
&= \text{Var}\left(\frac{\sum_1^n Y_i}{n+2}\right) \\
&= \frac{1}{(n+2)^2}\,\text{Var}\left(\sum_1^n Y_i\right) \\
&= \frac{np(1-p)}{(n+2)^2}
\end{aligned}$$

Then

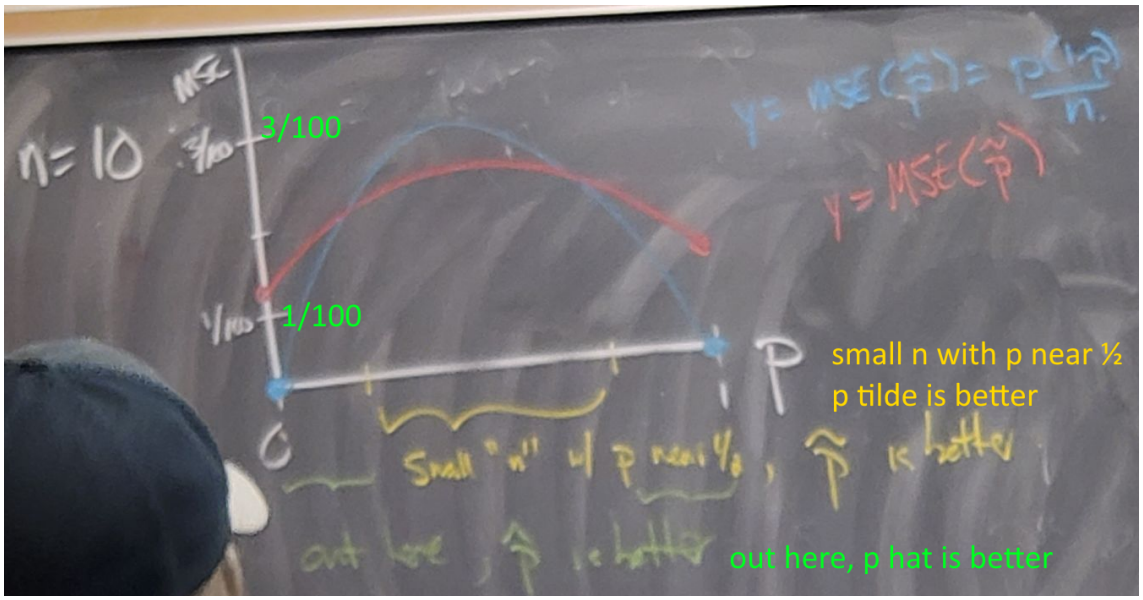
$$\begin{aligned}
\text{MSE}(\widetilde{p}) &= \text{Var}\,(\widetilde{p}) + \text{bias}(\widetilde{p})^2 \\
&= \frac{np(1-p)}{(n+2)^2} + \left(\frac{1-2p}{n+2}\right)^2 \\
&= \frac{np(1-p) + (1-2p)^2}{(n+2)^2}
\end{aligned}$$

Q: Which estimator is better? **A:** *it depends!*

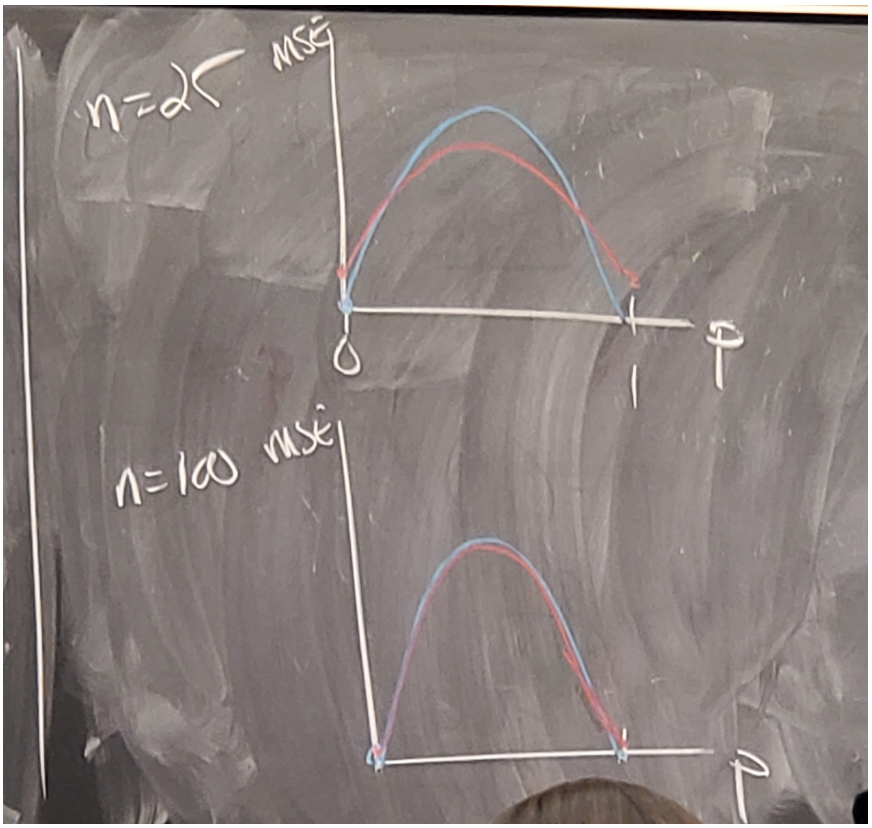
Note that for both, $\text{MSE}(\widetilde{p})(n,p)$ are functions of n and p. Let

$$y = \text{MSE}(\widetilde{p})(n,p) \qquad \text{and} \qquad y = \text{MSE}(\widehat{p})(n,p)$$

fix $n = 10$, $p \in [0, 1]$



$n = 25$ and 100 $p \in [0, 1]$



For larger n, they become indistinguishable.

8.3 Some Common Unbiased Point Estimators

Table 8.1, page 397: Common unbiased estimator for μ, ρ, $\mu_1 - \mu_2$, $\rho_1 - \rho_0$.

Example: Variance of a data set and the $n-1$.

Natural definition is $S^2 = \frac{1}{n}\sum \left(X_i - \bar{X}\right)^2$

Why do we use $n-1$?

Check bias (S^2). Is $\mathrm{E}\left(S^2\right) = \sigma^2$? Let

$$\mathrm{E}\left(X\right) = \mu \qquad \text{and} \qquad \mathrm{Var}\left(X\right) = \sigma^2$$

and $X_1, \ldots, X_n$ iid random sample. Then,

$$\begin{aligned}
\mathrm{E}\left(S^2\right) &= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right] \\
&= \frac{1}{n}\underbrace{\mathrm{E}\left[\sum\left(X_i^2 - 2X_i\bar{X} + \bar{X}^2\right)\right]}_{\text{focus on this}} \\
&= \frac{1}{n}\left(\mathrm{E}\left[\sum X_i^2 - 2\sum X_i\bar{X} + \sum \bar{X}^2\right]\right) \\
&= \frac{1}{n}\left(\underbrace{\mathrm{E}\left[\sum X_i^2\right]}_{①} - \underbrace{2\,\mathrm{E}\left[\sum X_i\bar{X}\right]}_{②} + \underbrace{\mathrm{E}\left[\sum \bar{X}\right]^2}_{③}\right)
\end{aligned}$$

① and ③ uses the same "trick" of $\mathrm{Var}\left(X\right) = \mathrm{E}\left[X^2\right] - \left(\mathrm{E}\left[X\right]\right)^2$. So,

$$\begin{aligned}
\text{Part } ① &= \mathrm{E}\left[\sum_{i=1}^{n} X_i^2\right] \\
&= \sum \mathrm{E}\left[X_i^2\right] \\
&= \sum\left(\mathrm{Var}\left(X_i\right) + \left(\mathrm{E}\left[X_i\right]\right)^2\right) \\
&= \sum\left(\sigma^2 + \mu^2\right) \\
&= n\left(\sigma^2 + \mu^2\right)
\end{aligned}$$

$$\begin{aligned}
\text{Part } ③ &= \mathrm{E}\left[\sum_{i=1}^{n} \bar{X}^2\right] \\
&= \sum\left(\mathrm{Var}\left(\bar{X}\right) + \left(\mathrm{E}\left[\bar{X}\right]\right)^2\right) \\
&= \sum\left(\frac{\sigma^2}{n} + \mu^2\right) \\
&= n\left(\frac{\sigma^2}{n} + \mu^2\right) \\
&= \sigma^2 + n\mu^2
\end{aligned}$$

$$
\begin{aligned}
\text{Part } \textcircled{2} &= -2\,\mathrm{E}\left[\sum_i X_i \left(\frac{1}{n}\sum_j X_j\right)\right] \\
&= -\frac{2}{n}\,\mathrm{E}\left[\sum_i \sum_j X_i X_j\right] \\
&= -\frac{2}{n}\,\mathrm{E}\left[\underbrace{\sum_i X_i X_j}_{n \text{ terms}} + \underbrace{\sum_{i\neq j}\sum X_i X_j}_{n^2-n \text{ terms}}\right] \\
&= -\frac{2}{n}\left(\underbrace{\sum_i \mathrm{E}\left[X_i^2\right]}_{\text{this is part } \textcircled{1} \text{ again}} + \underbrace{\sum_{i\neq j}\sum \mathrm{E}\left[X_i X_j\right]}_{\text{covariance like}}\right)
\end{aligned}
$$

Recall:

$$
\begin{aligned}
\mathrm{Cov}(X_i, X_j) &= \mathrm{E}\left[\left(X_i - \bar{X}\right)\left(X_j - \bar{X}\right)\right] \\
&= \mathrm{E}\left[X_i X_j\right] - \mathrm{E}\left[X_i\right]\mathrm{E}\left[X_j\right]
\end{aligned}
$$

But iid, "i" for independent $\implies \mathrm{Cov}(X_i, X_j) = 0$. Therefore,

$$
\mathrm{E}\left[X_i X_j\right] = \mathrm{E}\left[X_i\right]\mathrm{E}\left[X_j\right] = \mu \cdot \mu = \mu^2
$$

$$
\begin{aligned}
\text{Part } \textcircled{2} &= -\frac{2}{n}\Big(n(\sigma^2 + \mu^2) + (n^2 - n)\mu^2\Big) \\
&= -2\left[(\sigma^2 + \mu^2) + (n-1)\mu^2\right]
\end{aligned}
$$

Substituting back into the original parts,

$$
\begin{aligned}
\mathrm{E}\left(S^2\right) &= \frac{1}{n}\left(\underbrace{\mathrm{E}\left[\sum X_i^2\right]}_{①} - \underbrace{2\,\mathrm{E}\left[\sum X_i\bar{X}\right]}_{②} + \underbrace{\mathrm{E}\left[\sum \bar{X}\right]^2}_{③}\right) \\
&= \frac{1}{n}\left(\underbrace{n(\sigma^2+\mu^2)}_{①} - \underbrace{2\Big(\left(\sigma^2+\mu^2\right)+(n-1)\mu^2\Big)}_{②} + \underbrace{\sigma^2+n\mu^2}_{③}\right) \\
&= \frac{1}{n}\left(n\sigma^2+n\mu^2-2\sigma^2-2\mu^2-2(n-1)\mu^2+\sigma^2+n\mu^2\right) \\
&= \frac{1}{n}\left(n\sigma^2+n\mu^2-2\sigma^2-2\mu^2-2n\mu^2+2\mu^2+\sigma^2+n\mu^2\right) \\
&= \frac{1}{n}\left(n\sigma^2-2\sigma^2+\sigma^2\right) \\
&= \frac{1}{n}\left(n\sigma^2-\sigma^2\right) \\
&= \frac{\sigma^2(n-1)}{n} \\
&\neq \sigma^2
\end{aligned}
$$

Therefore, it is a biased point estimate.

To make an unbiased estimator, rescale the summation of the natural definition...

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

8.4 Evaluating the Goodness of a Point Estimator

Definition: Given any $\widehat{\theta}$, the exact (theoretical) error is defined as

$$\varepsilon = \left|\widehat{\theta} - \theta\right|$$

Note that ε is itself a random variable, and thus we can make probability statements about it. Obviously we want ε to be small. Consider $\left|\widehat{\theta} - \theta\right| < b$. We can make conclusions about

$$P\left(\left|\widehat{\theta} - \theta\right| < b\right)$$

Chebyshev:

$$P\left(\left|\widehat{\theta} - \theta\right| < k\sigma\right) > 1 - \frac{1}{k^2}$$

Remark: We call $\sigma = \sqrt{\sigma^2}$ the **standard error** in stats.

In practice, $k = 2$ is a common choice.

$$\Pr\left(\left|\widehat{\theta} - \theta\right| < 2\sigma\right) > 1 - \frac{1}{2^2} = \frac{3}{4} = 0.75$$

In the real world, 2σ is usually much better than this (see Table 8.2) when the underlying distribution $\widehat{\theta}$ is symmetric and unimodal.

Last working assumption for a few lectures: in real life we use $S^2 \approx \sigma^2$ (more later).

Example: (Difference in means)

2 iid random samples:

$$n_1 = 100 \qquad \bar{Y}_1 = 26{,}400 \qquad {S_1}^2 = 1{,}440{,}000$$

$$n_2 = 200 \qquad \bar{Y}_2 = 25{,}100 \qquad {S_2}^2 = 1{,}960{,}000$$

Estimating difference in means $\mu_1 - \mu_2$. Here,

$$\bar{Y}_1 - \bar{Y}_2 = \underbrace{26{,}400 - 25{,}100}_{\text{a point estimate}} = 1300$$

Recall: $\sigma^2_{\bar{Y}_1 - \bar{Y}_2} = \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}$. Use $S^2_{\bar{Y}_1 - \bar{Y}_2}$ to find a 2σ interval estimate:

$$\sigma^2_{\bar{Y}_1 - \bar{Y}_2} \approx S^2_{\bar{Y}_1 - \bar{Y}_2} = \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} \approx \frac{1{,}440{,}000}{100} + \frac{1{,}960{,}000}{200} = 22{,}800$$

$$\sigma_{\bar{Y}_1 - \bar{Y}_2} \approx S_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{S^2_{\bar{Y}_1 - \bar{Y}_2}} = \sqrt{22{,}800} \approx 151$$

Therefore, a 2σ interval estimate for $\mu_1 - \mu_2$ is $1300 \pm 2(151) \implies 1300 \pm 302$ or $(998, 1602)$

8.5 Confidence Intervals

Goal: Take interval estimates from §8.4 but make more precise comments about the probability $\Pr\left(\left|\widehat{\theta}-\theta\right|<k\sigma\right)$

$$-k\sigma<\widehat{\theta}-\theta<k\sigma \quad \Longleftrightarrow \quad \widehat{\theta}-k\sigma<\theta<\widehat{\theta}+k\sigma$$

Consider

$$\Pr\left(\widehat{\theta}_L\leq\theta\leq\widehat{\theta}_u\right)=1-\alpha$$

$\widehat{\theta}_L$ and $\widehat{\theta}_U$ are the lower and upper **confidence limits** , respectively.

$1-\alpha$ is the **confidence coefficient** .

Example:

Want 9% CI, chose $\alpha=5\%$ when we know how $\widehat{\theta}$ is distributed. We can use standardization methods to find the limits.

Two sided confidence interval:

$$\Pr\left(\widehat{\theta}_L\leq\theta\leq\widehat{\theta}_u\right)=1-\alpha$$

One sided:

$$\Pr\left(\widehat{\theta}_L\leq\theta\right)=1-\alpha \qquad \Pr\left(\widehat{\theta}_U\geq\theta\right)=1-\alpha$$

Discussion: The pivotal method for finding confidence intervals.

① We know how a r.v. Y is distributed, but not some underlying parameter θ.

② Using a distribution of an estimator $\widehat{\theta}$, we convert to a probability distribution that does not depend on θ (standardizing).

Example:

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

$$Z=\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}} \quad \Longrightarrow \quad \underbrace{N(0,1)}_{\text{Independent of }\mu}$$

We use $N(0,1)$ to rescale the limits of $\widehat{\theta}_L$ and $\widehat{\theta}_U$.

Example:

$Y_1,\ldots,Y_n$ an iid random sample of size n from a uniform distribution on the interval $(0,\ \theta)$.

We want an estimate for θ, use

$$\widehat{\theta}=\max\left\{Y_1,\ldots,Y_n\right\}$$

We know how $\widehat{\theta}$ is distributed, but it is clearly dependent on θ.

$$\text{For } Y_i\,, \quad f(y) = \frac{1}{\theta}$$

$$\implies F(y) = \Pr\left(Y \leq y\right) = \int_0^y \frac{1}{\theta} = \frac{y}{\theta}, \qquad y \in [0, \theta]$$

$$F(y) = \begin{cases} 0 & y < 0 \\ y/\theta & y \in [0, \theta] \\ 1 & y > \theta \end{cases}$$

The max order stat $\widehat{\theta}$ has CDF:

$$\begin{aligned} \Pr\left(\widehat{\theta} < w\right) &= \Pr\left(Y_1 \leq w,\ Y_2 \leq w,\ \ldots,\ Y_n \leq w\right) \\ &= \left[\Pr\left(Y \leq w\right)\right]^n \\ &= \begin{cases} 0 & w < 0 \\ (w/\theta)^n & w \in [0, \theta] \\ 1 & w > \theta \end{cases} \end{aligned}$$

Use a change of variables to find an associated pivotal distribution: $U = \dfrac{\widehat{\theta}}{\theta}$

The CDF of U,

$$\begin{aligned} \Pr\left(U \leq u\right) &= \Pr\left(\frac{\widehat{\theta}}{\theta} \leq u\right) \\ &= \Pr\left(\widehat{\theta} \leq \theta u\right) \\ &= \left(\frac{\theta u}{\theta}\right)^n \\ &= u^n \text{ for } u \in [0, 1] \end{aligned}$$

$$F(u) = \begin{cases} 0 & u < 0 \\ u^n & u \in [0, 1] \\ 1 & u > 1 \end{cases}$$

Pivotal CDF of u, no longer depends upon θ.

We use U's CDF to construct a confidence interval.

Goal: Find a 95% lower confidence interval for θ. Want:

$$\Pr\left(\underbrace{\widehat{\theta}_L \leq \theta}_{\text{One-sided C.I.}}\right) = 0.95$$

Using U's CDF: $\Pr(U \leq u) = 0.95 \iff u^n = 0.95$

$$u = \sqrt[n]{0.95} = (0.95)^{1/n}$$

Then,

$$\Pr\left(\frac{\widehat{\theta}}{\theta} \leq (0.95)^{1/n}\right) = 0.95$$

$$\Pr\left(\frac{\widehat{\theta}}{(0.95)^{1/n}} \leq \theta\right) = 0.95$$

But $\widehat{\theta} = \max(Y_1, \ldots, Y_n) = Y_{(n)}$

So, our 95% confidence interval is

$$\frac{Y_{(n)}}{(0.95)^{1/n}} \leq \theta.$$

Example:

For a random sample

$$\underbrace{0.76, 0.88, 1.68, 1.74, 1.78}_{n = 5}$$

$$Y_{(5)} = 1.78$$

A 95% C.I. for θ is given by $n = 5,\ Y_{(5)} = 1.78$

$$\frac{1.78}{(0.95)^{1/5}} \leq \theta$$

$$\frac{1.78}{0.98979} \leq \theta$$

$$1.798 \leq \theta$$

8.6 Large Sample Confidence Intervals

The unbiased point estimates for $\mu,\ \rho,\ \mu_1 - \mu_0,\ \rho_1 - \rho_2$ all have near Normal distributions by the Central Limit Theorem.

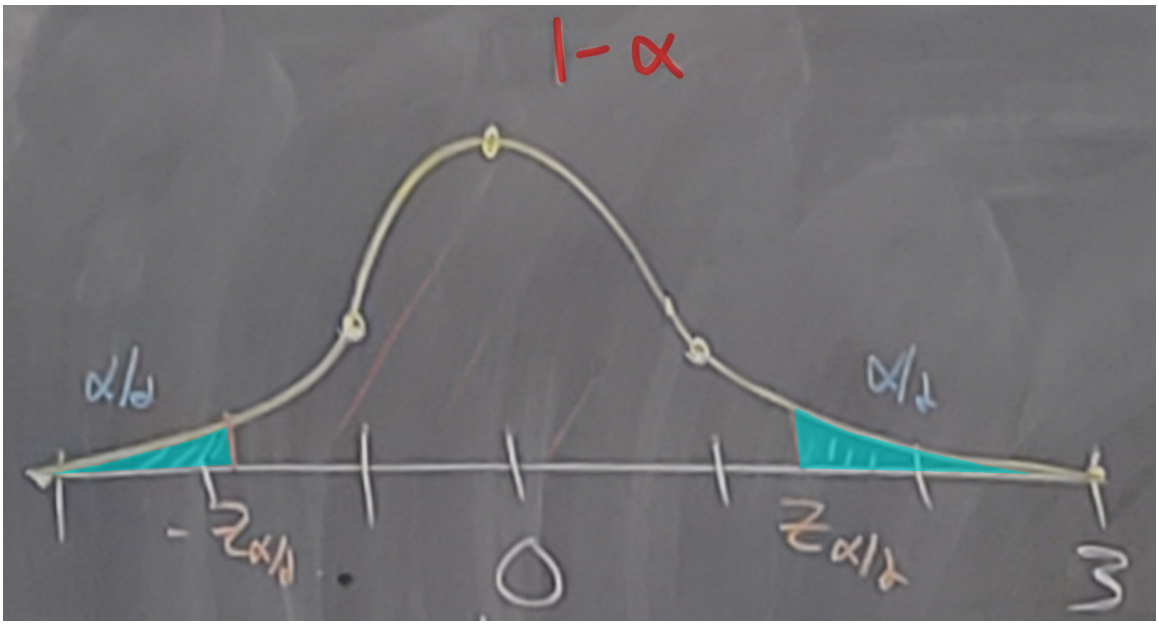
Moreover, using

$$Z = \frac{\widehat{\theta} - \theta}{\sigma_{\widehat{\theta}}}$$

Z is a pivotal quantity $N(0, 1)$.

For two-sided confidence intervals,

$$\Pr\left(-Z_{\alpha/2} \leq Z \leq Z_{\alpha/2}\right) = 1 - \alpha$$



Some common standard errors:

$$\begin{aligned} 90\%\text{ C.I.} &\implies Z_{0.05} = 1.645 \\ 95\%\text{ C.I.} &\implies Z_{0.025} = 1.960 \\ 99\%\text{ C.I.} &\implies Z_{0.005} = 2.576 \end{aligned}$$

Then

$$-Z_{\alpha/2} \leq Z \leq Z_{\alpha/2}$$

$$-Z_{\alpha/2} \leq \frac{\widehat{\theta} - \theta}{\sigma_{\widehat{\theta}}} \leq Z_{\alpha/2}$$

$$-Z_{\alpha/2}\sigma_{\widehat{\theta}} \leq \widehat{\theta} - \theta \leq Z_{\alpha/2}\sigma_{\widehat{\theta}}$$

$$-Z_{\alpha/2}\sigma_{\widehat{\theta}} - \widehat{\theta} \leq -\theta \leq Z_{\alpha/2}\sigma_{\widehat{\theta}} - \widehat{\theta}$$

$$\underbrace{\widehat{\theta} - Z_{\alpha/2}\sigma_{\widehat{\theta}}}_{\widehat{\theta}_L} \leq \theta \leq \underbrace{\widehat{\theta} + Z_{\alpha/2}\sigma_{\widehat{\theta}}}_{\widehat{\theta}_U}$$

Of course, we don't know $\sigma_{\widehat{\theta}}$ exactly. For "large" samples, we can use $\sigma_{\widehat{\theta}} \approx S_{\widehat{\theta}}$.

Example:

$$\bar{X} = 19.07 \qquad S^2 = 10.60 \qquad \text{with} \quad n = 32$$

Recall: For "large" samples ($n \geq 30$), $\bar{X}$ nearly distributed by $N(\mu, \sigma^2_{\bar{X}}) \approx N(\mu, \frac{10.60}{0.32})$. For a 95% CI for μ,

$$\sigma_{\bar{X}} \approx \sqrt{\frac{10.60}{0.32}} \approx 0.576$$

Here, $\alpha = 0.05$ and $\frac{\alpha}{2} = 0.025$. So $\bar{X} \pm Z_{0.025}\sigma_{\bar{X}}$, thus

$$19.07 \pm (1.96)(0.576)$$

$$\Rightarrow 19.07 \pm 1.128 \quad \text{or} \quad (17.94,\ 20.20)$$

Discussion: Differences in population proportions.

Estimating $p_1 - p_2$ by $\widehat{p}_1 - \widehat{p}_2$ for samples of size n_1 and n_2 respectively.

$$\widehat{\theta} \pm Z_{\alpha/2}\sigma_{\widehat{\theta}} \quad \Longrightarrow \quad \widehat{p}_1 - \widehat{p}_2 \pm Z_{\alpha/2}\sqrt{\sigma^2_{\widehat{p}_1} + \sigma^2_{\widehat{p}_2}}$$

$$\Longrightarrow \quad \widehat{p}_1 - \widehat{p}_2 \pm Z_{\alpha/2}\underbrace{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}}_{\substack{\text{Depends on the} \\ \text{unknowns } p_1,\, p_2}}$$

Two standard fixes:

① $pq = p(1-p) \leq 1/4$
Easy to show by calculus. Yields a "max" error via

$$\sqrt{\frac{1}{4n_1} + \frac{1}{4_n 2}}.$$

However, in practice we use ② as smaller confidence intervals are desirable.

② When n_i is large enough, we can use $\widehat{p}_1$ for p_i.

The $1 - \alpha$ confidence interval is

$$\widehat{p}_1 - \widehat{p}_2 \pm Z_{\alpha/2}\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}$$

8.7 Selecting the Sample Size

For our C.I. we have $\widehat{\theta} \pm Z_{\alpha/2}\sigma_{\widehat{\theta}}$

- $\sigma_{\widehat{\theta}}$ is the standards error
- $E = Z_{\alpha/2}\sigma_{\widehat{\theta}}$ is the error bounds of our C.I.

In real world, we choose our confidence level $1-\alpha$ and/or the error E ahead of time. This leads to the problem of choosing an appropriate sample size n.

Example: For $\bar{X}, \ldots$ distributions is $N(\mu, \sigma^2)$.

$$\sigma_{\bar{X}} = \sqrt{\sigma^2/n} \qquad \text{and} \qquad E = Z_{\alpha/2}\sqrt{\sigma^2/n}$$

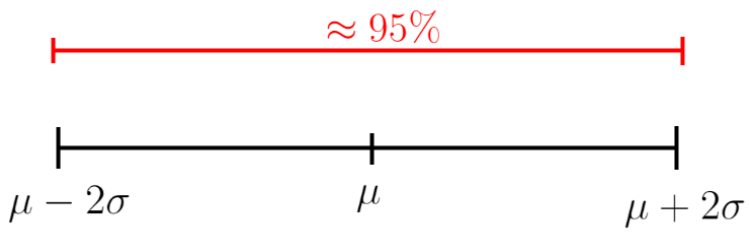
$$\longleftrightarrow \quad n = \frac{Z_{\alpha/2}\sigma^2}{E^2}$$

For desired E, we need to choose

$$n \geq \left\lceil \frac{Z_{\alpha/2}\sigma^2}{E^2} \right\rceil$$

Another issue: We don't really know $\sigma_{\widehat{\theta}}^2$ and maybe not even $Z_{\alpha/2}$.

① For $Z_{\alpha/2}$, we have seen that the empirical rule 2σ usually works for large n.



② For $\sigma_{\widehat{\theta}}$, either

a.) Use old data or a "pilot" sample S^2

b.) Use $\frac{1}{4}$ the spread of the data set, i.e.,

$$\sigma_{\widehat{\theta}} \approx \frac{1}{4}\big(\max(Y_i) - \min(Y_i)\big)$$

However, for proportions we can do even better.

Example: Proportions. For $\widehat{p} = \dfrac{\sum Y_i}{n}$ (relative frequency), $\widehat{p}$ distributed $N\left(p, \dfrac{pq}{n}\right)$

$$E = Z_{\alpha/2}\sqrt{\frac{pq}{n}} \qquad \longleftrightarrow \qquad n \geq \left\lceil \frac{Z_{\alpha/2}^2 p(1-p)}{E^2} \right\rceil$$

Again, since $p(1-p) \leq \dfrac{1}{4}$ on $[0, 1]$, we choose $n \geq \dfrac{Z_{\alpha/2}^2}{4E}$

Example: Public opinion polls ($\pm 3\%$)

For a 95% Confidence interval,

$$Z_{\alpha/2} = 1.960 \qquad \text{and} \qquad n \geq \left\lceil \frac{1.96^2}{4(0.03)^2} \right\rceil = 1068$$

For a 99% Confidence interval,

$$Z_{\alpha/2} = 2.576 \qquad \text{and} \qquad n \geq \left\lceil \frac{2.576^2}{4(0.03)^2} \right\rceil = 1844$$

8.8 Small-Sample Confidence Intervals for μ and $\mu_1 - \mu_2$

So for for $\bar{X}$ when n is large, we have used

$$\frac{\widehat{\theta} - \mu}{\sigma/\sqrt{n}} \approx \frac{\widehat{\theta} - \mu}{S/\sqrt{n}}$$

for small n, no longer precise enough.

But we set ourselves up for this issue back in Chapter 7.

Recall: We know $Z = \dfrac{\widehat{\theta} - \mu}{\sigma/\sqrt{n}}$ is distributed $N(0,\ 1)$

and $V = \dfrac{(n-1)S^2}{\sigma^2}$ is distributed χ^2 with $n-1$ df and Z and V are independent. Let

$$\begin{aligned} T &= \frac{\bar{X} - \mu}{S/\sqrt{n}} \\ &= \frac{\bar{X} - \mu}{S/\sqrt{n}} \cdot \frac{1}{s/\sigma} \\ &= \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \cdot \frac{1}{\sqrt{S^2/\sigma^2}} \\ &= Z \cdot \frac{1}{\sqrt{\dfrac{(n-1)S^2}{\sigma^2} \Big/ (n-1)}} \end{aligned}$$

We write $T = \dfrac{Z}{\sqrt{V/r}}$ where $r = (n-1)$.

T is the product of 2 pivotally distributed independent random variables.

Definition: We say T has a t-sampling distribution (t-dist'n) with $(n-1)$ degrees of freedom and its pdf is given by

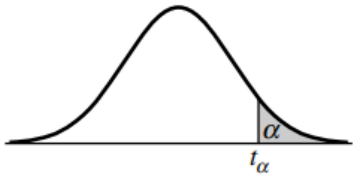
$$f_T(t) = \frac{\Gamma\left(\dfrac{r+1}{2}\right)}{\sqrt{\pi r} \cdot \Gamma\left(\dfrac{r}{2}\right)} \cdot \left(1 + \frac{t^2}{r}\right)^{-\frac{r+1}{2}}$$

Remarks:

① We won't work with $f_T(t)$ explicity. Note this distribution is pivotal... Independent of μ and σ^2. We pick our $t_{\alpha/2}$ from Table 5, page 849. Notice df $= r = n - 1 \geq 30$ are Z-scores:

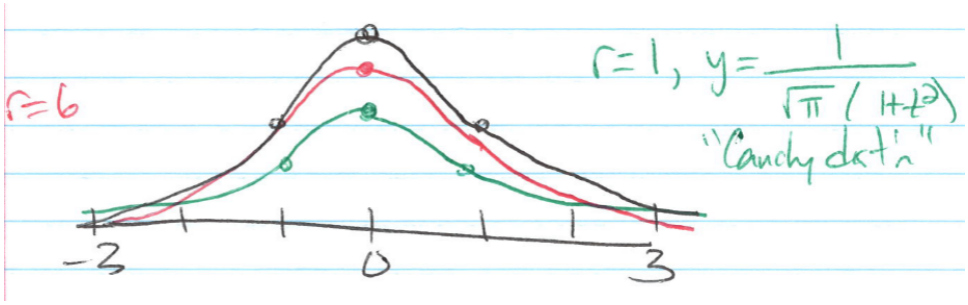
To do so we need α and d.f. $= n - 1$

Table 5 Percentage Points of the t Distributions



$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$	df
3.078	6.314	12.706	31.821	63.657	1
1.886	2.920	4.303	6.965	9.925	2
1.638	2.353	3.182	4.541	5.841	3
1.533	2.132	2.776	3.747	4.604	4
1.476	2.015	2.571	3.365	4.032	5
1.440	1.943	2.447	3.143	3.707	6
1.415	1.895	2.365	2.998	3.499	7
1.397	1.860	2.306	2.896	3.355	8
1.383	1.833	2.262	2.821	3.250	9
1.372	1.812	2.228	2.764	3.169	10
1.363	1.796	2.201	2.718	3.106	11
1.356	1.782	2.179	2.681	3.055	12
1.350	1.771	2.160	2.650	3.012	13
1.345	1.761	2.145	2.624	2.977	14
1.341	1.753	2.131	2.602	2.947	15
1.337	1.746	2.120	2.583	2.921	16
1.333	1.740	2.110	2.567	2.898	17
1.330	1.734	2.101	2.552	2.878	18
1.328	1.729	2.093	2.539	2.861	19
1.325	1.725	2.086	2.528	2.845	20
1.323	1.721	2.080	2.518	2.831	21
1.321	1.717	2.074	2.508	2.819	22
1.319	1.714	2.069	2.500	2.807	23
1.318	1.711	2.064	2.492	2.797	24
1.316	1.708	2.060	2.485	2.787	25
1.315	1.706	2.056	2.479	2.779	26
1.314	1.703	2.052	2.473	2.771	27
1.313	1.701	2.048	2.467	2.763	28
1.311	1.699	2.045	2.462	2.756	29
1.282	1.645	1.960	2.326	2.576	inf.

② t-distribution looks like "fat tailed" Normal Curves:



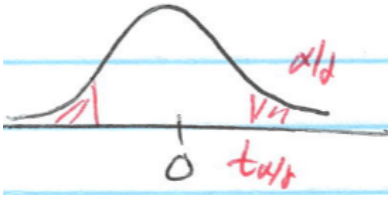
FACT: as $r \to \infty$, the D-distribution $\to N(0,\, 1)$.

FUN FACT: Cauchy distribution is "famous".

$$r = 1, \qquad \mu_T = 0, \quad \text{but} \quad \sigma_T^2 = \infty$$

Discussion: Small n Confidence Interval.

For 2 sided, $T = \dfrac{\bar{X} - \mu}{S/\sqrt{n}}$



$$\Pr\left(-t_{\alpha/2} \le T \le t_{\alpha/2}\right) = 1 - \alpha$$

$$\implies \bar{X} \pm t_{\alpha/2}\left(\frac{S}{\sqrt{n}}\right)$$

Example: $n = 10, \qquad \bar{X} = 3.22, \qquad S = 1.17$

Find a 95% confidence interval for μ.

Table 5: $\quad 95\% \longrightarrow t_{0.025}$ with df $= (n-1) = 10 - 1 = 9$.

Use $t_{0.025} = 2.262$ and then $\bar{X} \pm t_{0.025}\left(\dfrac{S}{\sqrt{n}}\right) \implies 3.22 \pm (2.262)\left(\dfrac{1.17}{\sqrt{10}}\right)$

$$3.22 \pm 0.84 \implies (2.38,\ 4.06)$$

BONUS MATH: From the t-distribution.

For the math curious, it comes down to a fancy change of variables.

Recall from last semester, of X, Y are independent, the joint pdf is $f(x,y) = f_x(x)f_y(y)$.

Here, $T = Z\left(\frac{V}{r}\right)^{-\frac{1}{2}}$ and Z, V are independent (as $\bar{X}$ and S are by Fisher's theorem).

To derive f_T, start with the join pdf of Z and Y, a χ^2 distribution of r degrees of freedom. Then,

$$f(y,z) = \frac{1}{\Gamma\left(\frac{r}{2}\right)2^{r/2}} \cdot y^{\frac{r}{2}-1} \cdot \exp\left[-\frac{r}{2}\right] \cdot \frac{1}{\sqrt{2\pi}} \cdot \exp\left[\frac{Z^2}{2}\right] \qquad y \in (0,\ \infty), \quad Z \in \mathbb{R}$$

Then let $t = Z\left(\frac{V}{r}\right)^{-\frac{1}{2}}$, $\qquad V = y$.

$$\begin{aligned}\Pr\left(Y < y,\ Z < z\right) &= \int_{-\infty}^{y}\int_{-\infty}^{z} f(y,z)\,dy\,dz \\ &= \int_{-\infty}^{T}\int_{-\infty}^{V} f\Big(y(t,v),\ z(t,v)\Big)\underbrace{\begin{vmatrix} Y_t & Y_v \\ Z_t & Z_v \end{vmatrix}^{-1}}_{\text{Jacobian}} dt\,dv\end{aligned}$$

Recall from MTH 420, for $u = h(x)$:

$$\int_a^b f(x)\,dx = \int_{h(a)}^{h(b)} f\left(h^{-1}(a)\right)\frac{d}{dx}\left(h^{-1}(u)\right)du$$

Discussion: Difference in means, small sample.

Start with X, $E[X] = \mu_x$, $\quad V[X] = \sigma_x^2$, (Normal).
And a r.v. Y, $E[Y] = \mu_y$, $\quad V[Y] = \sigma_y^2$, (Normal).

Now we take 2 different iid random samples, which yields

$$\bar{X} \text{ and } S_{\bar{X}}^2 \quad \text{with} \quad n_1$$

$$\bar{Y} \text{ and } S_{\bar{Y}}^2 \quad \text{with} \quad n_2$$

We have the unbiased estimator $\bar{X} - \bar{Y}$ with $V(\bar{X} - \bar{Y}) = \frac{\sigma_{\bar{X}}^2}{n_1} + \frac{\sigma_{\bar{Y}}^2}{n_2}$. Then,

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_{\bar{X}}^2}{n_1} + \frac{\sigma_{\bar{Y}}^2}{n_2}}} \quad \text{a pivotal quantity.}$$

Again, for small n_i, using $S_i \approx \sigma_i$ is not precise enough.

Assumption: Let $\sigma_{\bar{X}} = \sigma^{\bar{Y}}$. Thus

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \text{a pivotal quantity.}$$

Now we need to construct a point estimator for σ.

Idea: Try a "pooled" estimator.

$$S_p^2 = \frac{\sum(X_i - \bar{X})^2 + \sum(Y_i - \bar{Y})^2}{r}$$

$$S_p^2 = \frac{(n_1 - 1)S_{\bar{X}}^2 + (n_2 - 1)S_{\bar{Y}}^2}{r}$$

Claim: S_p^2 is unbiased when $r = \big[(n_1 - 1) + (n_2 - 1)\big] = n_1 + n_2 - 2$. (i.e. $E[\, S_p^2\,] = \sigma^2$)

Reason: It is really the same computation as when we showed that $E[\, S^2\,] = \sigma^2$ when $S = \frac{1}{n-1}\sum(X_i - \bar{X})^2$
Or, we can think about this as a product distribution of 2 independent χ^2 distributions. One with $n_1 - 1$ df and another with $n_2 - 1$ df. Then (like with mgfs) the product distribution is χ^2 with $n_1 - 1 + n_2 - 1$ df.

Definition: The **pooled estimator** is

$$S_p^2 = \frac{\sum(X_i - \bar{X})^2 + \sum(Y_i - \bar{Y})^2}{n_1 + n_2 - 2}.$$

This is unbiased, and now we can define the t-statistic for $\mu_X - \mu_Y$ as

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

Which has $n_1 + n_2 - 2$ degrees of freedom.

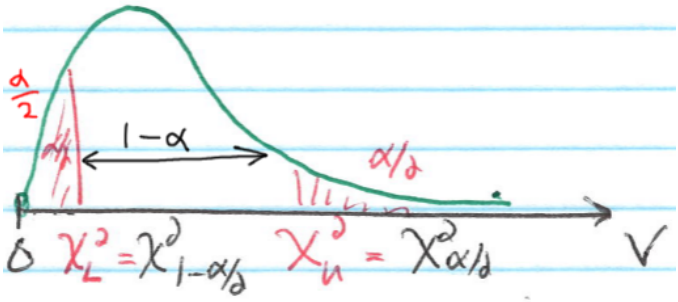
8.9 Confidence Intervals for σ^2

We have $S = \dfrac{1}{n-1}\sum(X_i - \bar{X})^2$ is unbiased and $V = \dfrac{(n-1)S^2}{\sigma^2}$ is a pivotal quantity with χ^2 distribution and $(n-1)$ df.

For a confidence interval, we use

$$\Pr\left(\chi^2_L \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_U\right) = 1 - \alpha$$

As with t-distributions, we need both our α and df $n-1$. But χ^2 itself is not symmetric. The most common C.I. makes symmetric probability in the "tails"



Then

$$\frac{\chi^2_L}{(n-1)S^2} \leq \frac{1}{\sigma^2} \leq \frac{\chi^2_U}{(n-1)S^2}$$

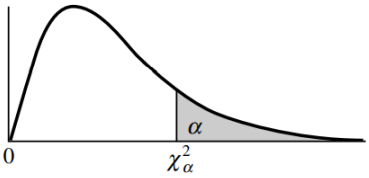
$$\implies \frac{1}{\chi^2_L} \geq \frac{\sigma^2}{(n-1)S^2} \geq \frac{1}{\chi^2_U}$$

$$\implies \frac{(n-1)S^2}{\chi^2_L} \geq \sigma^2 \geq \frac{(n-1)S^2}{\chi^2_U}$$

$$\iff \boxed{\frac{(n-1)S^2}{\chi^2_U} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_L}}$$

Confidence interval for σ^2

Table 6 Percentage Points of the χ^2 Distributions



df	$\chi^2_{0.995}$	$\chi^2_{0.990}$	$\chi^2_{0.975}$	$\chi^2_{0.950}$	$\chi^2_{0.900}$
1	0.0000393	0.0001571	0.0009821	0.0039321	0.0157908
2	0.0100251	0.0201007	0.0506356	0.102587	0.210720
3	0.0717212	0.114832	0.215795	0.351846	0.584375
4	0.206990	0.297110	0.484419	0.710721	1.063623
5	0.411740	0.554300	0.831211	1.145476	1.61031
6	0.675727	0.872085	1.237347	1.63539	2.20413
7	0.989265	1.239043	1.68987	2.16735	2.83311
8	1.344419	1.646482	2.17973	2.73264	3.48954
9	1.734926	2.087912	2.70039	3.32511	4.16816
10	2.15585	2.55821	3.24697	3.94030	4.86518
11	2.60321	3.05347	3.81575	4.57481	5.57779
12	3.07382	3.57056	4.40379	5.22603	6.30380
13	3.56503	4.10691	5.00874	5.89186	7.04150
14	4.07468	4.66043	5.62872	6.57063	7.78953
15	4.60094	5.22935	6.26214	7.26094	8.54675
16	5.14224	5.81221	6.90766	7.96164	9.31223
17	5.69724	6.40776	7.56418	8.67176	10.0852
18	6.26481	7.01491	8.23075	9.39046	10.8649
19	6.84398	7.63273	8.90655	10.1170	11.6509
20	7.43386	8.26040	9.59083	10.8508	12.4426
21	8.03366	8.89720	10.28293	11.5913	13.2396
22	8.64272	9.54249	10.9823	12.3380	14.0415
23	9.26042	10.19567	11.6885	13.0905	14.8479
24	9.88623	10.8564	12.4011	13.8484	15.6587
25	10.5197	11.5240	13.1197	14.6114	16.4734
26	11.1603	12.1981	13.8439	15.3791	17.2919
27	11.8076	12.8786	14.5733	16.1513	18.1138
28	12.4613	13.5648	15.3079	16.9279	18.9392
29	13.1211	14.2565	16.0471	17.7083	19.7677
30	13.7867	14.9535	16.7908	18.4926	20.5992
40	20.7065	22.1643	24.4331	26.5093	29.0505
50	27.9907	29.7067	32.3574	34.7642	37.6886
60	35.5346	37.4848	40.4817	43.1879	46.4589
70	43.2752	45.4418	48.7576	51.7393	55.3290
80	51.1720	53.5400	57.1532	60.3915	64.2778
90	59.1963	61.7541	65.6466	69.1260	73.2912
100	67.3276	70.0648	74.2219	77.9295	82.3581

Table 6 (***Continued***)

$\chi^2_{0.100}$	$\chi^2_{0.050}$	$\chi^2_{0.025}$	$\chi^2_{0.010}$	$\chi^2_{0.005}$	df
2.70554	3.84146	5.02389	6.63490	7.87944	1
4.60517	5.99147	7.37776	9.21034	10.5966	2
6.25139	7.81473	9.34840	11.3449	12.8381	3
7.77944	9.48773	11.1433	13.2767	14.8602	4
9.23635	11.0705	12.8325	15.0863	16.7496	5
10.6446	12.5916	14.4494	16.8119	18.5476	6
12.0170	14.0671	16.0128	18.4753	20.2777	7
13.3616	15.5073	17.5346	20.0902	21.9550	8
14.6837	16.9190	19.0228	21.6660	23.5893	9
15.9871	18.3070	20.4831	23.2093	25.1882	10
17.2750	19.6751	21.9200	24.7250	26.7569	11
18.5494	21.0261	23.3367	26.2170	28.2995	12
19.8119	22.3621	24.7356	27.6883	29.8194	13
21.0642	23.6848	26.1190	29.1413	31.3193	14
22.3072	24.9958	27.4884	30.5779	32.8013	15
23.5418	26.2962	28.8454	31.9999	34.2672	16
24.7690	27.5871	30.1910	33.4087	35.7185	17
25.9894	28.8693	31.5264	34.8053	37.1564	18
27.2036	30.1435	32.8523	36.1908	38.5822	19
28.4120	31.4104	34.1696	37.5662	39.9968	20
29.6151	32.6705	35.4789	38.9321	41.4010	21
30.8133	33.9244	36.7807	40.2894	42.7956	22
32.0069	35.1725	38.0757	41.6384	44.1813	23
33.1963	36.4151	39.3641	42.9798	45.5585	24
34.3816	37.6525	40.6465	44.3141	46.9278	25
35.5631	38.8852	41.9232	45.6417	48.2899	26
36.7412	40.1133	43.1944	46.9630	49.6449	27
37.9159	41.3372	44.4607	48.2782	50.9933	28
39.0875	42.5569	45.7222	49.5879	52.3356	29
40.2560	43.7729	46.9792	50.8922	53.6720	30
51.8050	55.7585	59.3417	63.6907	66.7659	40
63.1671	67.5048	71.4202	76.1539	79.4900	50
74.3970	79.0819	83.2976	88.3794	91.9517	60
85.5271	90.5312	95.0231	100.425	104.215	70
96.5782	101.879	106.629	112.329	116.321	80
107.565	113.145	118.136	124.116	128.299	90
118.498	124.342	129.561	135.807	140.169	100

Example: Construct a 90% C.I. for μ and σ^2.

$$\underbrace{85.4 \quad 86.8 \quad 86.1 \quad 85.3 \quad 84.8 \quad 86.0}_{\text{Data set, } n=6}$$

$$\bar{X} = 85.7\overline{3} \qquad S^2 = 0.502\overline{6} \qquad S \approx 0.7089$$

90% C.I. $\implies \alpha = 0.10,\ \dfrac{\alpha}{2} = 0.05$ with df $n - 1 = 5$.

$$\chi^2_L = \chi^2_{0.950} = 1.145476 \qquad \chi^2_U = \chi^2_{0.050} = 11.0705$$

A common mistake people make is forgetting that the lower bound is χ^2_U. For σ^2,

$$\frac{5 \cdot (0.502\overline{6})}{11.0705} \leq \sigma^2 \leq \frac{5 \cdot (0.502\overline{6})}{1.145475}$$

$$0.227 \leq \sigma^2 \leq 2.194$$

$$\sigma^2 \in (0.227,\ 2.194)$$

For μ, use a t-distribution with 5 degrees of freedom.

$$\bar{X} \pm t_{0.050} \cdot S \quad \implies \quad 85.7\overline{3} \pm (2.015)(0.7089)$$

$$\mu \in (84.305,\ 87.162)$$

Chapter 9

Properties of Point Estimators and Methods of Estimation

9.2 Relative Efficiency

In general, given two estimators $\widehat{\theta}_1$ and $\widehat{\theta}_2$ of a parameter θ, we claim the one with smaller MSE is better.

For unbiased estimators, MSE *is* variance (σ^2). So an idea of "better" is if $\text{Var}(\widehat{\theta}_1) < \text{Var}(\widehat{\theta}_2)$, we say $\widehat{\theta}_1$ is the better estimator.

Definition: Given unbiased estimators $\widehat{\theta}_1$ and $\widehat{\theta}_2$ of θ, the **efficiency** of $\widehat{\theta}_1$ relative to $\widehat{\theta}_2$ is defined by

$$\text{eff}(\widehat{\theta}_1,\ \widehat{\theta}_2) = \frac{\text{Var}(\widehat{\theta}_2)}{\text{Var}(\widehat{\theta}_1)}.$$

Note: If $\text{eff}(\widehat{\theta}_1,\ \widehat{\theta}_2) > 1$, then $\dfrac{\text{Var}(\widehat{\theta}_2)}{\text{Var}(\widehat{\theta}_1)} > 1 \implies \underbrace{\text{Var}(\widehat{\theta}_1) < \text{Var}(\widehat{\theta}_2)}_{\widehat{\theta}_1 \text{ is "better"}}$

Example: Let $Y_1, \ldots, Y_n$ be an iid random sample from $N(\mu,\ \sigma^2)$. Two estimators of σ^2 are:

$$\widehat{\sigma_1^2} = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

$$\widehat{\sigma_2^2} = \frac{1}{2}(Y_1 - Y_2)^2$$

We know S^2 is unbiased.

Show $\widehat{\sigma_2^2}$ is unbiased.

$$\begin{aligned}
\mathrm{E}(\widehat{\sigma_2^2}) &= \mathrm{E}\left(\frac{1}{2}\left(Y_1^2 - 2Y_1Y_2 + Y_2^2\right)\right) \\
&= \frac{1}{2}\Big[\mathrm{E}\left(Y_1^2\right) - 2\,\mathrm{E}\left(Y_1Y_2\right) + \mathrm{E}\left(Y_2^2\right)\Big] \\
&= \frac{1}{2}\Big[\mathrm{Var}\left(Y_1\right) + \big(\mathrm{E}\left(Y_1\right)\big)^2 - 2\,\mathrm{E}\left(Y_1\right)\mathrm{E}\left(Y_2\right) + \mathrm{Var}\left(Y_2\right) + \big(\mathrm{E}\left(Y_2\right)\big)^2\Big] \\
&= \frac{1}{2}\Big[\sigma^2 + \mu^2 - 2\mu\cdot\mu + \sigma^2 + \mu^2\Big] \\
&= \sigma^2 \qquad \text{unbiased.}
\end{aligned}$$

To compute relative efficiency, we need variances.

For $\mathrm{Var}\left(S^2\right)$, we know $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$.

By properties of χ^2 distribution, $\mathrm{Var}\left(\dfrac{(n-1)S^2}{\sigma^2}\right) = 2(n-1)$.

Then, $\dfrac{(n-1)^2}{\sigma^4}\cdot\mathrm{Var}\left(S^2\right) = 2(n-1) \quad\Longrightarrow\quad \mathrm{Var}\left(S^2\right) = \dfrac{2\sigma^4}{n-1}$.

For $\mathrm{Var}(\widehat{\sigma_2^2})$, this takes more work.

$$\mathrm{Var}\left(\frac{1}{2}(Y_1-Y_2)^2\right) = \underbrace{\mathrm{E}\left(\left(\frac{1}{2}(Y_1-Y_2)^2\right)^2\right)}_{*} - \underbrace{\left[\mathrm{E}\left(\frac{1}{2}(Y_1-Y_2)^2\right)\right]^2}_{\sigma^4}$$

$$\begin{aligned}
* &= \frac{1}{4}\,\mathrm{E}\left((Y_1-Y_2)^4\right) \\
&= \frac{1}{4}\,\mathrm{E}\left(Y_1^4 - 4Y_1^3Y_2 + 6Y_1^2Y_2^2 - 4Y_1Y_2^3 + Y_2^4\right) \\
&= \frac{1}{4}\Big[\mathrm{E}\left(Y_1^4\right) - 4\,\mathrm{E}\left(Y_1^3\right)\mathrm{E}\left(Y_2\right) + 6\,\mathrm{E}\left(Y_1^2\right)\mathrm{E}\left(Y_2^2\right) - 4\,\mathrm{E}\left(Y_1\right)\mathrm{E}\left(Y_2^3\right) + \mathrm{E}\left(Y_2^4\right)\Big] \\
&= \frac{1}{4}\Big[2\,\mathrm{E}\left(Y^4\right) - 8\,\mathrm{E}\left(Y^3\right)\mathrm{E}\left(Y\right) + 6\,\mathrm{E}\left(Y^2\right)^2\Big]
\end{aligned}$$

We need the higher moments m_3' and m_4'. Recall the mgf for $N(\mu, \sigma^2)$ is $m(t) = \exp\left[\mu t + \dfrac{t^2\sigma^2}{2}\right]$

$$
\begin{aligned}
m'(t) &= (\mu + t\sigma^2)m(t) \\
m'(0) &= \mu m(0) \\
&= \mu \\
&= \mathrm{E}\,(Y)
\end{aligned}
$$

$$
\begin{aligned}
m''(t) &= \sigma^2 m(t) + (\mu + t\sigma^2)^2 m(t) \\
m''(0) &= \sigma^2 + \mu^2 m(0) \\
&= \sigma^2 + \mu^2 \\
&= \mathrm{E}\left(Y^2\right)
\end{aligned}
$$

$$
\begin{aligned}
m^{(3)}(t) &= \sigma^2 m'(t) + 2(\mu + t\sigma^2)\sigma^2 m(t) + (\mu + t\sigma^2)^2 m'(t) \\
m^{(3)}(0) &= \sigma^2 m'(0) + 2(\mu + 0\sigma^2)\sigma^2 m(0) + (\mu + 0\sigma^2)^2 m'(0) \\
&= \sigma^2\mu + 2\mu\sigma^2 + \mu^3 \\
&= \mu^3 + 3\mu\sigma^2 \\
&= \mathrm{E}\left(Y^3\right)
\end{aligned}
$$

$$
\begin{aligned}
m^{(4)}(t) &= \cdots \\
m^{(4)}(0) &= \cdots \\
&= \mu^4 + 6\mu^2\sigma^2 + 3\sigma^4 \\
&= \mathrm{E}\left(Y^4\right)
\end{aligned}
$$

Substituting back into $*$, we get $* = \dfrac{1}{4}(12\sigma^4) = 3\sigma^4$. Thus,

$$
\mathrm{Var}\left(\widehat{\sigma_2^2}\right) = \mathrm{Var}\left(\frac{1}{2}(Y_1 - Y_2)^2\right) = \underbrace{3\sigma^4}_{*} - \underbrace{\sigma^4}_{\sigma^4} = 2\sigma^4
$$

Hence, $\mathrm{eff}(S^2, \widehat{\sigma_2^2}) = \dfrac{2\sigma^4}{\frac{2\sigma^4}{n-1}} = n - 1$

9.3 Consistency

Discussion: "Covergence in probability"

This is a different type of analysis than that of Calculus.

Example: MTH 420, "pointwise" convergence:

$$f_n(x) = x^n, \quad x \in [0,\, 1] \qquad \lim_{n\to\infty} f_n(x) = \begin{cases} 0 & x \in [0,\, 1) \\ 1 & x = 1 \end{cases}$$

Let $x \in (0, 1)$. To show $x^n \to 0$:
For any $\epsilon > 0$, need an index N such that when $n > N$ we get $|x^n - 0| < \epsilon$

$$\text{Choose} \quad N > \left\lceil \frac{\ln \epsilon}{\ln x} \right\rceil$$

In probability, convergence is about the probabalistic measure of an event.

Example: Law of Large Numbers:

A, an event associated with an event E. $\Pr(A) = p$.

Do n iid repetitions of E and count $n_A :=$ number of times A occurs.

The relative frequency $f_A = \dfrac{n_A}{n}$. The Law of Large Numbers says $f_A \to p$

What does this mean exactly? Conclusion: $\underbrace{\Pr(|f_A - p| < \epsilon)}_{*} \geq 1 - \dfrac{p(1-p)}{n\epsilon^2}$

* This means the probability that p is in the confidence interval $f_A - \epsilon \leq p \leq f_A + \epsilon$.

Convergence? $\lim_{n\to\infty} \Pr(|f_A - p| < \epsilon) \geq \lim_{n\to\infty} \left(1 - \dfrac{p(1-p)}{n\epsilon^2}\right) = 1$

$$\implies \lim_{n\to\infty} \Pr(|f_A - p| < \epsilon) = 1$$

This is convergence in probability!

In probability, convergence is about the probabalistic measure of an event (measure theory).

Definition: Estimatior $\widehat{\theta}$ is **consistent** if $\widehat{\theta} \to \theta$ in probability.

That is $\widehat{\theta}_n$ is a consistent estimator of θ if for any $\epsilon > 0$

$$\lim_{n\to\infty} \Pr\left(\left|\widehat{\theta}_n - \theta\right| \leq \epsilon\right) = 1$$

$$\text{or} \quad \lim_{n\to\infty} \Pr\left(\left|\widehat{\theta}_n - \theta\right| > \epsilon\right) = 0$$

Discussion: A tool for showing consistency:

Theorem An unbiased estimator $\widehat{\theta}_n$ for θ is consistent if

$$\lim_{n\to\infty} \text{Var}(\widehat{\theta}_n) = 0$$

Proof. Let $\epsilon > 0$, then consider,

$$\Pr\left(\left|\widehat{\theta}_n - \theta\right| > \epsilon\right)$$

Recall Chebyshev's Theorem

$$\Pr(|Y - \mu| > k\sigma) \leq \frac{1}{k^2}$$

Here $\epsilon = k\sigma_{\widehat{\theta}_n} \implies k = \dfrac{\epsilon}{\sigma_{\widehat{\theta}_n}}, \qquad 0 \leq \Pr\left(\left|\widehat{\theta}_n - \theta\right| > \epsilon\right) \leq \dfrac{1}{(\epsilon/\sigma_{\widehat{\theta}_n})^2}$

Then

$$\Pr\left(\left|\widehat{\theta}_n - \theta\right| > \epsilon\right) \leq \frac{\sigma^2_{\widehat{\theta}_n}}{\epsilon^2} = \frac{\text{Var}(\widehat{\theta}_n)}{\epsilon^2}$$

As $n \to \infty$,

$$0 \leq \lim_{n\to\infty} \Pr\left(\left|\widehat{\theta}_n - \theta\right| > \epsilon\right) \leq \lim_{n\to\infty} \frac{\text{Var}(\widehat{\theta}_n)}{\epsilon^2} = 0$$
$$\implies \Pr\left(\left|\widehat{\theta}_n - \theta\right| > \epsilon\right) = 0$$

Hence, $\widehat{\theta}_n \to \theta$ in probability. □

Example: (Common) Consistent Estimators

① $\bar{Y}$ know unbiased for μ and

$$\text{Var}(\bar{Y}) = \frac{\sigma^2}{n}$$

Provided that σ^2 is finite, $\lim_{n\to\infty} \text{Var}(\bar{Y}) = 0$

$\bar{Y}$ is consistant, that is, $\bar{Y} \to \mu$

② $\widehat{p}$ unbiased for p.

$$\text{Var}\left(\widehat{p}\right) = \frac{pq}{n} \quad \text{and} \quad \lim_{n\to\infty} \text{Var}\left(\widehat{p}\right) = 0$$
$$\implies \widehat{p} \to p \text{ i.e. consistent}$$

Example: (9.17 & 9.18)
$X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ are iid random samples with μ_x and μ_y but $\text{Var}(X) = \text{Var}(Y) = \sigma^2$.

a.) Show that $\bar{X} - \bar{Y}$ is a consistent estimator of $\mu_x - \mu_y$.

Solution: We showed in Chapter 8 that $\text{E}(\bar{X} - \bar{Y}) = \mu_x - \mu_y$ is unbiased.

Then $\text{Var}(\bar{X} - \bar{Y}) = \text{Var}(\bar{X}) + (-1)^2 \text{Var}\left(\bar{Y}\right) = \dfrac{2\sigma^2}{n}$.

Because $\lim_{n\to\infty} \text{Var}(\bar{X} - \bar{Y}) = 0$, we have consistency.

b.) Show that the pooled estimator,

$$S_p^2 = \frac{\sum_1^n (X_i - \bar{X})^2 + \sum_1^n (Y_i - \bar{Y})^2}{2n-2}$$

is a consistent estimator of σ^2 when $X, Y \sim N(\mu, \sigma^2)$.

Solution: To show unbiased we need Var $\left(S_p^2\right)$.

"Trick," convert to a distribution we understand well.

$$\underbrace{\frac{(2n-2)S_p^2}{\sigma^2}}_{\substack{\chi^2(2n-2) \\ \text{By Fisher's}}} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 + \sum_{i=1}^n (Y_i - \bar{Y})^2}{\sigma^2}$$

$$= \underbrace{\sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sigma}\right)^2 + \sum_{i=1}^n \left(\frac{Y_i - \bar{Y}}{\sigma}\right)^2}_{U_n \sim \chi^2(2n-2)}$$

$$\mathrm{E}\left(\frac{(2n-2)S_p^2}{\sigma^2}\right) = 2n-2 \text{ by properties of } \chi^2$$

$$\frac{2n-2}{\sigma^2}\,\mathrm{E}\left(S_p^2\right) = 2n-2 \implies \mathrm{E}\left(S_p^2\right) = \sigma^2 \quad \text{unbiased.}$$

For consistency, use same "trick."

$$\mathrm{Var}\left(\frac{(2n-2)S_p^2}{\sigma^2}\right) = \mathrm{Var}\left(U_n\right)$$

$$\left(\frac{2n-2}{\sigma^2}\right)^2 \mathrm{Var}\left(S_p^2\right) = 2 \cdot (2n-2) \quad \text{by properties of } \chi^2$$

$$\mathrm{Var}\left(S_p^2\right) = \frac{2\sigma^4}{2n-2}$$

$$\lim_{n\to\infty} \mathrm{Var}\left(S_p^2\right) = 0 \text{ therefore } S_p^2 \text{ is consistent.}$$

Theorem: "The Limit Laws" (For probability convergence)

Let $\widehat{\theta}_n \to \theta$ and $\widehat{\Psi}_n \to \Psi$. Then,

① $\widehat{\theta}_n + \widehat{\Psi}_n \to \theta + \Psi$

② $\widehat{\theta}_n \cdot \widehat{\Psi}_n \to \theta \cdot \Psi$

③ $\dfrac{\widehat{\theta}_n}{\widehat{\Psi}_n} \to \dfrac{\theta}{\Psi}$

④ If g is a continuous function at θ, then $g(\widehat{\theta}_n) \to g(\theta)$

Example: S_p^2 again.

$$\begin{aligned} S_p^2 &= \frac{\sum_1^n (X_i - \bar{X})^2 + \sum_1^n (Y_i - \bar{Y})^2}{2n-2} \\ &= \frac{(n-1)S_{\bar{X}}^2 + (n-1)S_{\bar{Y}}^2}{2n-2} \\ &= \frac{S_{\bar{X}}^2 + S_{\bar{Y}}^2}{2} \end{aligned}$$

But we already know $S_{\bar{X}}^2 \to \sigma^2$ and $S_{\bar{Y}}^2 \to \sigma^2$

Hence, by the Limit laws,

$$S_p^2 \to \frac{\sigma^2 + \sigma^2}{2} = \sigma^2$$

Discussion: Consistency and large n confidence intervals

We have $\bar{Y}_n \pm Z_{\alpha/2}\dfrac{S_n}{\sqrt{n}}$ By consistency, $\bar{Y}_n \to \mu$ and $S_n \to \sigma$

Hence, when σ is finite, $\bar{Y}_n \pm Z_{\alpha/2}\frac{S_n}{\sqrt{n}} \to \mu$

Hence $\lim\limits_{n\to\infty} \Pr\left(\left|\bar{Y}_{(n)} - \mu < Z_{\alpha/2}\dfrac{S_n}{\sqrt{n}}\right| < \epsilon\right) = 1$

That is, $\bar{Y}_{(n)} \pm Z_{\alpha/2}\dfrac{S_n}{\sqrt{n}} \to \mu$ in probability.

"Consistency justified our working assumption in Chapter 8"

Example: (9.24)

Let $Z_1, \ldots, Z_n$ be an iid random sample with $Z \sim N(0,1)$

Let $U_n = \sum_{i=1}^{n} Z_i^2 \qquad Z = \dfrac{Z_i - \mu}{\sigma}$

We "know" $U_n \sim \chi^2(n)$

Showed in old HW that $Z^2 \sim X^2(n)$

In notes, we used the product of n mgfs to show that $U_n \sim \chi^2(n)$.

Hence, $\mathrm{E}(U_n) = n$ and $\mathrm{Var}(U_n) = 2n$. Let

$$W_n := \frac{1}{n} U_n$$

Note that $\mathrm{E}(W_n) = \dfrac{1}{n}\mathrm{E}(U_n) = 1$

Hence, W_n is an unbiased estimator when $\sigma^2 = 1$

$$\begin{aligned}
\mathrm{Var}(W_n) &= \mathrm{Var}\left(\frac{1}{n} U_n\right) \\
&= \frac{1}{n^2}\mathrm{Var}(U_n) \\
&= \frac{1}{n^2} \cdot 2n \\
&= \frac{2}{n}
\end{aligned}$$

Since $\lim_{n\to\infty} \mathrm{Var}(W_n) = 0$, we get $W_n \to 1$ in probability.

9.4 Sufficiency

Discussion: Among a collection of estimators, how do we choose which one to work with and why?

Conceptually: A statistic is deemed **sufficient** if it "contains all of the available information about a parameter," given some random sample.

Example:

Statistician 1 has the data $X_1, \ldots, X_n$ and computes an estimator $\widehat{\theta}$.

Statistician 2 has only the stat $T = \tau(X_1, \ldots, X_n)$ estimating θ.

Sufficient implies both statisticians can make equally correct estimates about θ. Equivalently, a sufficient estimator $\widehat{\theta}$ utilizes all of the information in a sample relating to θ.

In either case, both are able to make the "same" conclusions about θ.

Definition: Let $Y_1, \ldots, Y_n$ be a sample from a probability distribution that is known up to parameter θ.

The statistic $T = \tau(Y_1, \ldots, Y_n)$ is **sufficient for θ** if the conditional distribution of $Y_1, \ldots, Y_n$ given T does not depend on θ.

Remark: **Sufficiency Principle:**

Given 2 data sets $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ where the stats

$$\tau(X_1, \ldots, X_n) = \tau(Y_1, \ldots, Y_n) = T$$

Any inference about θ should be the same regardless of the data set.

Example: (One time only via the definition)

Let $Y \sim \text{Bern}(p)$
i.e. $\Pr(Y = 1) = p$, $\Pr(Y = 0) = 1 - p$.

Let Y_n be a sequence of observations. We have $\widehat{p} = \dfrac{\sum Y_i}{n}$.

Sufficiency Principle:

Any sequence with k 1's and $n - k$ 0's is going to yield the same $\widehat{p}$. In other words, same inferences about p.

Let $S = \sum Y_i$

To show S is sufficient for p, we must show that the conditional distribution of the data $Y_1, \ldots, Y_n$ given S does not depend on p.

For $k \in (0, 1, 2, \ldots, n)$, $S = Y_1 + \cdots + Y_n = k$. We have,

$$\Pr\left(Y_1 = y_1, \ldots, Y_n = y_n \mid S = k\right) = \frac{\Pr\left(Y_1 = y_1, \ \ldots, \ Y_n = y_n \text{ and } S = k\right)}{\Pr(S = k)}$$

$$= \begin{cases} 0 & \sum Y_i \neq k \\ \dfrac{\Pr\left(Y_1 = y_2, \ \ldots, \ Y_n = y_n \text{ and } S = k\right)}{\Pr(S = k)} & \sum Y_i = k \end{cases}$$

$$= \frac{\overbrace{\prod_{i=1}^{n} \Pr(Y_i = y_i)}^{\text{independent}}}{\underbrace{\Pr(S = k)}_{\substack{\text{Binomial, } k \text{ ``successes''} \\ \text{in } n \text{ trials}}}}$$

$$= \frac{p^{y_1}(1-p)^{1-y_1} \times \cdots \times p^{y_n}(1-p)^{1-y_n}}{\binom{n}{k} p^k (1-p)^{n-k}}$$

$$= \frac{p^{\sum_1^n Y_i}(1-p)^{n - \sum_1^n Y_i}}{\binom{n}{k} p^k (1-p)^{n-k}}$$

$$= \frac{p^k (1-p)^{n-k}}{\binom{n}{k} p^k (1-p)^{n-k}}$$

$$= \begin{cases} 0 & S \neq k \\ \dfrac{1}{\binom{n}{k}} & S = k \quad \leftarrow \text{ independent of } p \end{cases}$$

By definition, $S = \sum Y_i$ is sufficient.

Remark: Started with $\widehat{p} = \frac{S}{n}$.

Note: Any one-to-one transformation (injection) of a sufficient stat will again be sufficient.

Example: $f(x) = \dfrac{x}{n}$ is an injection (has invertible form).

$$S \text{ sufficient} \implies \frac{S}{n} \text{ sufficient}$$

Thus $\widehat{p}$ is also sufficient.

Example: $\sqrt{x}$ is also an injection on [0, 1].

Thus $\sqrt{S}$ is sufficient.

Topic: The Factorization Theorem:

In practice we do not use the definition.

Theorem: The Factorization Theorem

Let $X_1, \ldots, X_n$ denote a random sample from a distribution with pdf $f(x \mid \theta)$

(distribution depends on the unknown parameter θ)

The statistic $T = \tau(x_1, \ldots, x_n)$ is a sufficient statistic for θ if and only if the conditional pdf can be factored as follows.

$$L(x_1, \ldots, x_n \mid \theta) = g(\tau(x_1, \ldots, x_n) \mid \theta) \cdot h(x_1, \ldots, x_n)$$

Where

- g is a function that depends on the data only through the stat T
- h does not depend on θ

Remarks:

① Via independence (of data)

$$L(x_1, \ldots, x_n \mid \theta) = f(x_1 \mid \theta) \cdot f(x_2 \mid \theta) \times \cdots \times f(x_n \mid \theta)$$

The product of the marginals.

② The function L is called the **likelihood of the sample.**

Example: Let $X_1, \ldots, X_n$ denote a random sample from a Poisson distribution with parameter $\lambda > 0$. Find a sufficient statistic for the parameter λ.

Here,

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!}\ ,\ x = 0,\ 1,\ 2,\ \ldots$$

$$\begin{aligned} L(x_1,\ x_2\ \ldots,\ x_m \mid \lambda) &= f(x_1 \mid \lambda) \cdot f(x_2 \mid \lambda) \times \cdots \times f(x_n \mid \lambda) \\ &= \frac{\lambda^{x_1} e^{-\lambda}}{x_1!} \cdot \frac{\lambda^{x_2} e^{-\lambda}}{x_2!} \times \cdots \times \frac{\lambda^{x_n} e^{-\lambda}}{x_n!} \\ &= \underbrace{\lambda^{x_1 + \cdots + x_n} e^{-\lambda n}}_{\substack{S := \sum X_i \\ g(S \mid \lambda)}} \cdot \underbrace{\frac{1}{x_1! \cdot x_2! \cdot \cdots \cdot x_n!}}_{h(x_1, \ldots, x_n)} \end{aligned}$$

With $S = \sum X_i$,

$$g(S \mid \lambda) = e^{-\lambda n} \lambda^S$$

and

$$h(x_1, \ldots, x_n) = \frac{1}{x_1! \cdot x_2! \cdot \cdots \cdot x_n!}$$

By Factorization Theorem, S is a sufficient statistic.

Note: $\bar{X} = \frac{1}{n}\sum X_i$ or $n\bar{X} = S = \sum X_i$. So,

$$g(S \mid \lambda) = e^{-\lambda}\lambda^{n\bar{X}} = g(\bar{X} \mid \lambda)$$

Therefore $\bar{X}$ is also a sufficient statistic.

Example: The $\widehat{p}$ example again. $Y \sim \text{Bern}(p)$.

$$\begin{aligned} L(y_1, \ldots, y_n \mid p) &= f(y_1 \mid p) \times \cdots \times f(y_n \mid p) \\ &= p^{y_1}(1-p)^{1-y_1} \times \cdots \times p^{y_n}(1-p)^{1-y_n} \\ &= p^{\sum y_i}(1-p)^{n-\sum y_i} \qquad S := \sum y_i \\ &= \underbrace{p^S(1-p)^{n-S}}_{g(S \mid p)} \cdot \underbrace{1}_{h(\vec{y})} \end{aligned}$$

Where $\vec{y}$ is the n tuple $y_1, y_2, \ldots, y_n$.

By Factorization Theorem, S is sufficient.

Yesterday: Sufficiency $\underbrace{f(x_1, \ldots, x_n \mid S)}_{S \text{ is sufficient for } \theta} = \underbrace{\dfrac{f(x_1, \ldots, x_n \text{ and } S(p) = s)}{f(s)}}_{\text{conditional prob is independent of } \theta}$

Factorization theorem:

$$L(x_1, \ldots, x_n \mid \mu) = \underbrace{g(\mu,\ s)}_{\substack{S \text{ is} \\ \text{sufficient} \\ \text{for } \theta}} \cdot h(\vec{x})$$

Example: Let $X_1, \ldots, X_n$ be a random sample from $N(\mu, 1)$. (unknown μ, $\sigma^2 = 1$.)

$$f(x) = \frac{1}{\sqrt{2\pi}\exp\left[\dfrac{-(x-\mu)^2}{2}\right]}$$

Find a sufficient statistic for μ.

Construct Likelihood function,

$$\begin{aligned}
L(x_1, \ldots, x_n \mid \mu) &= f(x_1 \mid \mu) \cdots f(x_n \mid \mu) \\
&= \frac{1}{\sqrt{2\pi}} \exp\left[\frac{-(x_1-\mu)^2}{2}\right] \times \cdots \times \frac{1}{\sqrt{2\pi}} \exp\left[\frac{-(x_n-\mu)^2}{2}\right] \\
&= \left(\frac{1}{\sqrt{2\pi}}\right)^2 \exp\left[-\frac{1}{2}\sum (x_i-\mu)^2\right] \\
&= \frac{1}{(2\pi)^{n/2}} \exp\left[-\frac{1}{2}\sum (x_i^2 - 2x_i\mu + \mu^2)\right] \\
&= \frac{1}{(2\pi)^{n/2}} \exp\left[-\frac{1}{2}\sum (-2x_i\mu + \mu^2)\right] \cdot \exp\left[-\frac{1}{2}\sum x_i^2\right] \\
&= \frac{1}{(2\pi)^{n/2}} \exp\left[-\frac{1}{2}\left(-2\mu\sum x_i + n\mu^2\right)\right] \cdot \exp\left[-\frac{1}{2}\sum x_i^2\right] \\
&= \frac{1}{(2\pi)^{n/2}} \underbrace{\exp\left[\mu\sum x_i - \frac{1}{2}n\mu^2\right]}_{\text{Define } S := \sum x_i} \exp\left[-\frac{1}{2}\sum x_i^2\right]
\end{aligned}$$

$$g(S \mid \mu) = \frac{1}{(2\pi)^{n/2}} \exp\left[\mu S - \frac{n\mu^2}{2}\right]$$

$$h(\vec{x}) = \exp\left[-\frac{1}{2}\sum x_i^2\right]$$

By Factorization Theorem, S is a sufficient statistic for μ.

Example: (9.49)

Suppose $X_i \sim \text{Unif}(0,\, \theta)$, θ unknown. Then,

$$f(x) = \frac{1}{\theta}\,, \qquad x \in [0,\, \theta]$$

(In this, we use $X_{(n)} = \max\{x_1, \ldots, x_n\}$ as an estimator of θ.)

$f(x_1 \mid \theta) \times \cdots \times f(x_n \mid \theta) = \dfrac{1}{\theta^n}$ independent of any x_i's.

To bring the x_i's into the story, we use the indicator function

$$I_{[0,\,\theta]}(x) = \begin{cases} 1 & x \in [0, \theta] \\ 0 & \text{else} \end{cases}$$

So,

$$f(x) = \frac{I_{[0,\,\theta]}(x)}{\theta}$$

With this,

$$\begin{aligned}
L(x_1, \ldots, x_n \mid \theta) &= \frac{I_{[0,\,\theta]}(x_1)}{\theta} \times \cdots \times \frac{I_{[0,\,\theta]}(x_n)}{\theta} \\
&= \frac{1}{\theta^n} I_{[0,\,\theta]}(x_i \in [0, \theta],\ i = 1, 2, \ldots, n)
\end{aligned}$$

Now, $x_i \leq \theta$ for all i if and only if $\max\{x_1, \ldots, x_n\} \leq 0$. So,

$$L(x_1, \ldots, x_n \mid \theta) = \frac{1}{\theta^n} I_{[0,\,\theta]}(\max\{x_1, \ldots, x_n \leq \theta\})$$

Define statistic $T := \max\{x_1, \ldots, x_n\}$ and then

$$g(T \mid \theta) = \frac{1}{\theta^n} I_{[0,\,\theta]}(T)$$

$$h(x_1, \ldots, x_n) = 1$$

By the Factorization Theorem, $T = \max(x_1, \ldots, x_n)$ is a sufficient statistic.

Note: $\bar{X}$ is not sufficient since there is no way to bring $\bar{X}$ into L.
(Of course, no reason we should be able to do so since $\bar{X}$ is also biased for $\theta \ldots$ and clearly not consistent. $\bar{X} \to \dfrac{\theta}{2}$ and these $\bar{X} \to \theta$ in probability)

9.5 Rao-Blackwell Theorem and Minimum-Variance Unbiased Estimation

Theorem: The Rao-Blackwell Theorem

Let X and Y be a random variable such that $\mathrm{E}(Y) = \mu$ and $\mathrm{Var}(Y) = \sigma^2 y$.

Let $\phi(x) = \mathrm{E}(Y \mid X)$.

Then $\mathrm{E}(\phi(x)) = \mu$ and $\sigma^2_{\phi(x)} \leq \sigma^2 y$

Proof. Have $f(x, y)$, $f_1(x)$, $f_2(y)$, and

$$h(y \mid x) = \frac{f(x, y)}{f_1(x)}*$$

$$\begin{aligned}
\phi(x) &= \mathrm{E}(\phi(x)) \\
&= \int_{\mathbb{R}} y h(y \mid x) dy. \\
&= \int_{\mathbb{R}} y \frac{f(x, y)}{f_1(x)} dy \\
&= \frac{1}{f_1(x)} \int_{\mathbb{R}} y f(x, y) dy
\end{aligned}$$

Then

$$f_1(x)\phi(x) = \int_{\mathbb{R}} y f(x, y) dy$$

$$\begin{aligned}
\mathrm{E}(\phi(x)) &= \int_{\mathbb{R}} \phi(x) f_1(x) dx \\
&= \int_{\mathbb{R}} \left[\int_{\mathbb{R}} y f(x, y) dy \right] dx \\
&= \int_{\mathbb{R}} \int_{\mathbb{R}} f(x, y) dy dx \\
&= \int_{\mathbb{R}} y \left[\int_{\mathbb{R}} f(x, y) dx \right] dy \\
&= \int_{\mathbb{R}} y f_2(y) dy \\
&= \mathrm{E}(Y) \\
&= \mu.
\end{aligned}$$

Note that $\sigma^2_{\phi(x)} = \mathrm{E}\left((\phi(x)-\mu)^2\right)$.

$$\begin{aligned}
\sigma_y^2 &= \mathrm{E}\left((y-\mu)^2\right) \\
&= \mathrm{E}\left(((y-\phi(x))(\phi(x)-\mu)^2\right) \\
&= \mathrm{E}\left((y-\phi(x))^2 + 2(y-\phi(x))(\phi(x)-\mu) + (\phi(x)-\mu)^2\right) \\
&= \underbrace{\mathrm{E}\left((y-\phi(x))^2\right)}_{\geq\ 0} + \underbrace{2\,\mathrm{E}\left((y-\phi(x))(\phi(x)-\mu)\right)}_{*} + \underbrace{\mathrm{E}\left((\phi-\mu)^2\right)}_{\sigma^2_{\phi(x)}} \\
(*) &= \int_{\mathbb{R}}\int_{\mathbb{R}} (y-\phi(x))(\phi(x)-\mu) f(x,y)\,dydx \\
&= \int_{\mathbb{R}} (\phi(x)-\mu)\left[\int_{\mathbb{R}} (y-\phi(x)) f(x,y)\,dy\right]dx \\
(*) &= \int_{\mathbb{R}} (\phi(x)-\mu)\left[\int_{\mathbb{R}} (y-\phi(x)) f_1(x) h(y\mid x)\,dy\right]dx \\
&= \int_{\mathbb{R}} (\phi(x)-\mu) f_1(x)\left[\underbrace{\int_{\mathbb{R}} (y-\phi(x)) h(y\mid x)\,dy}_{\text{note}^{**}}\right]dx \\
&= 0
\end{aligned}$$

Note**: This is zero as

$$\phi = \mathrm{E}\left(y \mid x\right) = \int y h(y\mid x)\,dy$$

So $\sigma_y^2 = \mathrm{E}\left((y-\phi(x))^2\right) + \sigma^2_{\phi(x)} \geq \sigma^2_{\sigma(x)}$

Aside: This almost always a strict inequality. For equality,

$$\Pr\left(Y-\phi=0\right) = 0$$

□

Discussion: Using the RBT (Rao-Blackwell Theorem) to construct "best" statistics.

Big Idea: We can take any unbiased estimator of θ and combine with a sufficient stat for θ to get a "better" estimator.

Let $x_1,\ldots,x_n$ denote a random sample of $f(x\mid\theta)$. We have $T=\tau(x_1,\ldots,x_n)$ sufficient for θ. Let $U=u(x_1,\ldots,x_n)$ be an unbiased stat for θ which is not itself a function of T. Consider $\mathrm{E}\left(u\mid T\right)$:
Since T is sufficient, the conditional probability of U given $T=\tau$ does not depend upon θ.

Define a new stat

$$\phi(\tau) = \mathrm{E}\left(U\mid T\right)$$

which is a function of τ alone. Hence, $\phi(T)$ is a statistic (does not depend opon θ). (i.e. only a function of the data.)

By RBT, $\phi(T)$ is an unbiased statistic for θ with the guarantee that

$$\sigma^2_{\phi(T)} < \sigma^2_u$$

Theorem: Rao-Blackwell Theorem (Textbook version):
Let $\widehat{\theta}$ be an unbiased estimator for θ such that $\text{Var}(\widehat{\theta}) < \infty$.
If T is a sufficient statistic for θ, define $\widehat{\theta}^* = \text{E}\left(\widehat{\theta}^* \mid T\right)$.

Then, for all θ, $\text{E}\left(\widehat{\theta}^*\right) = \theta$ and

$$\text{Var}(\widehat{\theta}^*) \leq \text{Var}(\widehat{\theta})$$

Discussion: Minimum Variance Unbiased Estimator (MVUE)

If $\widehat{\theta}$ is an MVUE, then

$$\text{Var}(\widehat{\theta}) \leq \text{Var}(\widehat{\Psi}) \quad \text{for any other estimator } \widehat{\Psi}.$$

Typically (almost always) the process we described above lends to an MVUE.

① $\widehat{\theta}$ unbiased estimator for θ

② T sufficient for θ

③ $\text{E}\left(\widehat{\theta} \mid T\right)$ is a MVUE

Remark: All of yesterdays sufficiency examples indicate that all traditional estimators are MVUE.

Example: Bern(p)
Here unbiased estimator $\widehat{p} = \dfrac{S}{n}$.
Here $S = \sum Y_i$ is sufficient.

$$\text{E}\left(\widehat{p} \mid S\right) = \frac{S}{n} \text{ sufficient stat, has no } p.$$

Rao-Blackwell $\implies$ $\widehat{p}$ is an MVUE.

Also note, $\text{Var}\left(\widehat{p}\right) < \text{Var}\left(S\right)$, is much better $\dfrac{p}{n} < np$.

Example: Last day we showd that $S = \sum Y_i$ is a sufficient stat for λ in Pois(λ).

$$f(y) = \frac{\lambda^y e^{\lambda}}{y!}\,, \quad y = 0, 1, 2, \ldots$$

We seek an MVUE of $e^{-\lambda}$. Need an unbiased estimator of $Y_1, \ldots, Y_n$ iid.

a) Show that

$$W = \begin{cases} 1 & Y_1 = 0 \\ 0 & Y_1 \neq 0 \end{cases}$$

is unbiased for $e^{-\lambda}$. (New idea: Not just looking for λ, but a function of λ.)

Solution: $\mathrm{E}(W) = 1 \cdot \Pr(Y_1 = 0) + 0 \cdot \Pr(Y_1 \neq 0) = \Pr(Y_1 = 0) = \dfrac{\lambda^0 e^{-\lambda}}{0!} = e^{-\lambda}$

b) Use RBT to find a MVUE. Compute $\mathrm{E}(W \mid S)$

$$\begin{aligned}
\mathrm{E}(W \mid S) &= 1 \cdot \Pr(Y_1 = 0 \mid Y_1 + Y_2 + \cdots + Y_n = S) + 0 \cdot \Pr(Y_1 \neq 0 \mid S) \\
&= \frac{\Pr(Y_1 = 0 \text{ and } Y_1 + \cdots + Y_n = S)}{\Pr(Y_1 + \cdots + Y_n = S)} \\
&= \frac{\Pr(Y_1 = 0 \text{ and } Y_2 + \cdots + Y_n = S)}{\Pr(Y_1 + \cdots + Y_n = S)} \quad \text{top now independent events} \\
&= \frac{\Pr(Y_1 = 0) \cdot \Pr(Y_2 + \cdots + Y_n = S)}{\Pr(Y_1 + \cdots + Y_n = S)} \\
&= \frac{e^{-\lambda} \cdot \Pr(Y_2 + \cdots Y_n = s)}{\Pr(Y_1 + \cdots + Y_n = s)}
\end{aligned}$$

325 FACT: How is $Y_1 + \cdots + Y_n$ distributed? $Y_i \sim \mathrm{Pois}(\lambda)$.

Via moment generating functions, we proved that $Y_1 + \cdots + Y_n \sim \mathrm{Pois}(n\lambda)$

Hence $Y_2 + \cdots + Y_n \sim \mathrm{Pois}((n-1)\lambda)$, so

$$\begin{aligned}
\mathrm{E}(W \mid S) &= \frac{e^{-\lambda}\left(\dfrac{((n-1)\lambda)^S e^{-(n-1)\lambda}}{S!}\right)}{\dfrac{(n\lambda)^S e^{-n\lambda}}{S!}} \\
&= e^{-\lambda}\left(\frac{(n-1)\lambda}{n\lambda}\right)^S \cdot \frac{e^{-(n-1)\lambda}}{e^{-n\lambda}} \\
&= e^{-\lambda} \cdot \left(\frac{n-1}{n}\right)^S e^{\lambda} \\
&= \left(\frac{n-1}{n}\right)^S
\end{aligned}$$

By RBT, $\phi(S) = \left(\dfrac{n-1}{n}\right)^S$ is an MVUE of $e^{-\lambda}$, $S = \sum Y_i$.

Aside:

$$e^{-\lambda} = \lim_{n\to\infty}\left(1 - \frac{1}{n}\right)^n$$

$$\phi(S) = \lim_{n\to\infty}\left(1 - \frac{1}{n}\right)^S$$

9.6 The Method of Moments

This is out oldest estimation technique. If there are k parameters that have to be estimated, set the 1^{st} k population moments (given in terms of the parameters) and set them equal to the sample moments.

Recall:

$$\underbrace{\mu_k' = \mathrm{E}\left(Y^k\right)}_{\text{population}} \qquad \text{and} \qquad \underbrace{m_k' = \frac{1}{n}\sum Y_i^k}_{\text{data set}}$$

Example: Let $Y \sim N(\mu, \sigma^2)$. Then

$$\mu_1' = \mathrm{E}\left(Y\right) = \mu$$
$$\mu_2' = \mathrm{E}\left(Y^2\right) = \sigma^2 + \mu^2$$

$$m_1' = \frac{1}{n}\sum Y_i = \bar{Y}$$
$$m_2' = \frac{1}{n}\sum Y_i^2$$

Set equal and solve for paramters, then

$$\mu = \bar{Y}$$
$$\sigma^2 + \mu^2 = \frac{1}{n}\sum Y_i^2$$

Want σ^2 in terms of sample moments.

$$\sigma^2 + \bar{Y}^2 = \frac{1}{n}\sum Y_i^2$$
$$\sigma^2 = \frac{1}{n}\sum Y_i^2 - (\bar{Y})^2$$

Yields 2 estimators:

$$\widehat{\theta} = \bar{Y} \quad \text{for } \mu$$
$$\widehat{\Psi} = \frac{1}{n}\sum Y_i^2 - (\bar{Y})^2 \quad \text{for } \sigma^2$$

Note: $\widehat{\Psi}$ is our "old" biased estimator for σ^2. While both of these are consistent stats, $\widehat{\Psi}$ is biased as $\mathrm{E}(\widehat{\Psi}) = \dfrac{n-1}{n}\sigma^2$.

Example: Suppose $Y_1, \ldots, Y_n \sim \text{Gamma}\,(\alpha, \beta)$. Use MoM to find estimators for α and β.

Recall for Gamma, $\mathrm{E}\left(Y\right) = \alpha\beta$, $\mathrm{Var}\left(Y\right) = \alpha\beta^2$. So,

$$
\begin{aligned}
\mu_1' &= \alpha\beta \\
m_1' &= \bar{Y} \\
\mu_2' &= \frac{1}{n}\sum Y_1^2 \\
&= \mathrm{E}\left(Y^2\right) \\
&= \mathrm{Var}\left(Y\right) + \left(\mathrm{E}\left(Y\right)\right)^2 \\
&= \alpha\beta^2 + (\alpha\beta)^2 \\
&= \beta^2(\alpha + \alpha^2) \\
m_2' &= \frac{1}{n}\sum Y_i^2
\end{aligned}
$$

System: $\alpha\beta = \bar{Y}$ $\beta^2(\alpha + \alpha^2) = m_2'$ $\alpha = \dfrac{\bar{Y}}{\beta}$ Then,

$$
\begin{aligned}
m_2' &= \beta^2\left(\frac{\bar{Y}}{\beta} + \frac{\bar{Y}^2}{\beta^2}\right) \\
&= \bar{Y}^2 + \beta\bar{Y} \\
\iff \quad \beta\bar{Y} &= m_2' - \bar{Y}^2 \\
\iff \quad \beta &= \frac{m_2' - \bar{Y}^2}{\bar{Y}} \\
\implies \quad \alpha &= \frac{\bar{Y}^2}{m_2' - \bar{Y}^2}
\end{aligned}
$$

Define α's estimator $\widehat{\theta} := \dfrac{\bar{Y}^2}{m_2' - \bar{Y}^2}$.

and β's estimator $\widehat{\Psi} := \dfrac{m_2' - \bar{Y}^2}{\bar{Y}}$.

The following is the last example using MoM:

Example: $Y_1, \ldots, Y_n \sim \mathrm{Unif}(0,\, \theta)$, with θ unknown.

$$
\mu_1' = \mathrm{E}\left(Y\right) = \frac{\theta}{2} \qquad\qquad m_1' = \frac{1}{n}\sum Y_i = \bar{Y}
$$

Mom set equal solution for population parameters.

$$
\frac{\theta}{2} = \bar{Y} \quad \implies \quad \theta = 2\bar{Y}
$$

Define $\widehat{\theta} := 2\bar{Y}$.

So this is another estimator for θ. Is it any good?

Well, it is unbiased since $\mathrm{E}\left(2\bar{Y}\right) = 2 \cdot \frac{\theta}{2} = \theta$ and it is consistent.

Use variance trick to show...

$$\operatorname{Var}\left(2\bar{Y}\right) = 4\operatorname{Var}\left(\bar{Y}\right) = 4\frac{\operatorname{Var}(Y)}{n} = 4 \cdot \frac{\theta^2/12}{n} = \frac{\theta^2}{3n}$$

Note:

$$\lim_{n\to\infty} \operatorname{Var}\left(2\bar{Y}\right) = \lim_{n\to\infty} \frac{\theta^2}{3n} = 0$$

Consistent by theorem. However, earlier we showed $\widehat{\Psi} = \max\{Y_1, \ldots, Y_n\}$ is sufficient. Consider $\operatorname{Var}(\widehat{\Psi})$.

Like in Homework, $F(y) = \displaystyle\int_0^y \frac{1}{\theta}\, dy = \frac{y}{\theta}$

$$\Pr\left(Y_{(n)} \le y\right) = \left(\frac{y}{\theta}\right)^n = \frac{y^n}{\theta^n}.$$

And $f_{(n)}(y) = \dfrac{ny^{n-1}}{\theta^n}$, thus

$$\mathrm{E}(\widehat{\Psi}) = \int_0^\theta y \cdot f_{(n)}(y)\, dy = \frac{n}{n+1}\theta.$$

$\widehat{\Psi}$ is biased but it is consistent since $\lim\limits_{n\to\infty} \mathrm{E}[\,\widehat{\Psi}\,] \to \theta$. For $\operatorname{Var}(\widehat{\Psi}) = \mathrm{E}(\widehat{\Psi}^2) - \mathrm{E}(\widehat{\Psi})^2$:

$$\mathrm{E}(\widehat{\Psi}^2) = \int_0^\theta y^2 \cdot \frac{ny^{n-1}}{\theta^n} = \frac{n}{n+2}\theta^2$$

So...

$$\operatorname{Var}(\widehat{\Psi}) = \frac{n}{n+2}\theta^2 - \left(\frac{n}{n+1}\theta\right)^2 = \frac{n\theta^2}{(n+1)^2(n+2)}$$

And this variance is much better than $\operatorname{Var}(\widehat{\theta})$.

$$\operatorname{bias}(\widehat{\theta}) = 0$$

$$\operatorname{bias}(\widehat{\Psi}) = \frac{n}{n+1}\theta - \theta = -\frac{\theta}{n+1}$$

$$\operatorname{MSE}(\widehat{\theta}) = \operatorname{Var}\left(\widehat{\theta}\right) + \operatorname{bias}(\widehat{\theta})^2 = \frac{\theta^2}{3n}$$

$$\operatorname{MSE}(\widehat{\Psi}) = \frac{n\theta^2}{(n-2)(n+1)^2} + \frac{\theta^2}{(n+1)^2}$$

We see both $\operatorname{Var}(\widehat{\Psi}) < \operatorname{Var}(\widehat{\theta})$ and $\operatorname{MSE}(\widehat{\Psi}) < \operatorname{MSE}(\widehat{\theta})$

Of course, if we were to define

$$\widetilde{\Psi} := \frac{n+1}{n}\widehat{\Psi}$$

Now we have an unbiased consistent statistic. Last day, we showed that $\widehat{\Psi}$ is sufficient. So by the Rao-Blackwell Theorem, $\widetilde{\Psi}$ will be an MVUE. That is, $\operatorname{Var}(\widetilde{\Psi}) \le \operatorname{Var}(\widehat{\Psi})$.

9.7 The Method of Maximum Likelihood

Recall, the likelihood function is simply the joint pdf with the interpretation that a parameter(s) is unknown.

$$Y_1 \sim \text{ via } f(y \mid \theta)$$

Y_i's are iid random sample.

If θ is known, the joint pdf is

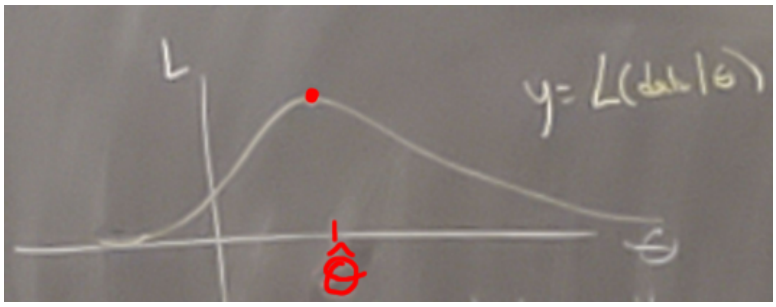
$$\begin{aligned} \underbrace{f(y_1, \ldots, y_n)}_{\text{domain } f \subseteq \mathbb{R}^n} &= \underbrace{f_1(y_1) \cdot f_2(y_2) \cdots f_n(y_n)}_{\text{product of marginals}} && \text{by independence} \\ &= f(y_1) \cdot f(y_2) \cdots f(y_n) && \text{by i.i.d.} \\ &= \prod_{i=1}^{n} f(y_i) \end{aligned}$$

If θ unknown, we emphasize this via the **likelihood function** .

$$\underbrace{L(Y_1, \ldots, Y_n \mid \theta)}_{\text{domain } L \subseteq \mathbb{R}^{n+1}} = \underbrace{f(y_1 \mid \theta) \cdot f(y_2 \mid \theta) \times \cdots \times f(y_n \mid \theta)}_{\text{really the same}}$$

Discussion: The reason this is called a "likelihood" function. Given a set of observations Y_i, what is the most likely value of θ.

Fix $(Y_1, \ldots, Y_n)$, L at this number is a 1 variable function in terms of θ.



Given this data set, what is the most likely value that θ takes? That would be θ corresponding to the largest value of pdf $\prod_i^n f(y_i \mid \theta)$.

How do we find the largest? (Optimize with respect to θ). That is, set $\dfrac{\partial L}{\partial \theta} = 0$ solve and verify.

Definition: The solution to $L_\theta = 0$ defines an estimator $\widehat{\theta}$ of θ, called the maximum likelihood estimator (MLE).

Remark: In computations, verify the critical point *is* a maximum.

Example: $X \sim \text{Bern}(p) = \text{Binom}(1, p)$

Given $X_1, \ldots, X_n$ an iid sample.

$$\begin{aligned} L(X_1, \ldots, X_n \mid p) &= p^{X_1}(1-p)^{1-X_1} \cdots p^{X_n}(1-p)^{1-X_n} \\ &= p^S(1-p)^{n-S}, \qquad S := \sum X_i \end{aligned}$$

Then,

$$\begin{aligned} \frac{\partial L}{\partial p} &= Sp^{S-1}(1-p)^{n-S} + p^S(n-s)(1-p)^{n-S-1}(-1) \\ &= p^{S-1}(1-p)^{n-S-1}\left[S(1-p) - p(n-S)\right] \end{aligned}$$

And $\dfrac{\partial L}{\partial p} = 0$ when

$$\begin{aligned} S - Sp - pn + Sp &= 0 \\ S - pn &= 0 \\ p &= \frac{S}{n} = \bar{X} \end{aligned}$$

Is this a max? Yes, use the 1st derivitive test: $\dfrac{\partial L}{\partial p} = p^{S-1}(1-p)^{n-S-1}\left[S - pn\right]$.

If $p < \dfrac{S}{n}$, then $\dfrac{\partial L}{\partial p} > 0$.

If $p > \dfrac{S}{n}$, then $\dfrac{\partial L}{\partial p} < 0$

Therefore $\bar{X}$ is an MLE for p.

Discussion: The log-likelihood function.

$$L(\vec{X} \mid \theta) = \prod_{i=1}^{n} f(X_i \mid \theta)$$

Always an n-product. Depending on f, computing $\frac{\partial L}{\partial \theta}$ can be difficult.

Since $f(X_i \mid \theta) > 0$ and we have a function whose existance is to turn products into sums...

Definition: The log-likelihood function:

$$\ln L(\vec{X} \mid \theta)$$

Claim: The MLE of $L(\vec{X} \mid \theta)$ also maximizes $\ln L(\vec{X} \mid \theta)$

Reason:

$$\frac{\partial}{\partial \theta}\left(\ln L(\vec{X} \mid \theta)\right) = \frac{\dfrac{\partial L}{\partial \theta}\left(\vec{X} \mid \theta\right)}{L(\vec{X} \mid \theta)} = 0 \qquad \Longleftrightarrow \qquad \frac{\partial}{\partial \theta}\left(L(\vec{X} \mid \theta)\right) = 0$$

Example: The last one again...

$$L(X_1, \ldots, X_n \mid p) = p^S (1-p)^{n-S}, \quad S := \sum X_i$$

$$\ln L(\vec{x} \mid p) = S \ln p + (n - S) \ln(1 - p)$$

$$\frac{\partial}{\partial p} (\ln L) = \frac{S}{p} + (n - s) \cdot \frac{-1}{1-p}$$

$$\begin{aligned} \frac{\partial}{\partial p} (\ln L) = 0 \quad &\iff \quad S(1-p) + (n-S)(-p) = 0 \\ &\iff \quad S - Sp - np + Sp = 0 \\ &\iff \quad p = \frac{S}{n} \end{aligned}$$

Topic: More population parameters.

This idea scales. If f depends upon k unknowns $\theta_1, \ldots, \theta_k$, then define

$$L(X_1, \ldots, X_n \mid \theta_1, \ldots, \theta_k) = \prod_{i=1}^{n} f(x_i \mid \theta_1, \ldots, \theta_k)$$

Optimizing this is a calculus problem.

Remark: We could have also extended the idea of the Factorization Theorem the same way with sufficient statistics to get

$$L(x_1, \ldots, x_n \mid \theta_1, \ldots, \theta_k) = g(S_1, \ldots, S_k, \theta_1, \ldots, \theta_k) \cdot h(x_1, \ldots, x_n)$$

Example: Let $X_1, \ldots, X_n$ be an iid from $N(\mu, \sigma^2)$ with both parameters unknown. Find the MLE for μ and σ^2.

For ease of computation, $N(\mu, \theta)$ such that $\theta := \sigma^2$.

$$L(x_1, \ldots, x_n \mid \mu_1 = \theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\theta}} \exp\left[-\frac{(X_i - \mu)^2}{2\theta} \right]$$

Fixing $\vec{x}$, we consider $L_{\vec{x}}(\mu,\ \theta)$.... I have no intentions of $\nabla L_{\vec{x}}$ under a product sign...

$$\begin{aligned} \ln L &= \sum_{i=1}^{n} \left(\ln\left(\frac{1}{\sqrt{2\pi\theta}} \right) + \frac{-(X_i - \mu)^2}{2\theta} \right) \\ &= -\frac{1}{2} \ln(2\pi\theta) n - \frac{1}{2\theta} \sum_{i=1}^{n} (X_i - \mu)^2 \end{aligned}$$

$$\begin{aligned} \frac{\partial}{\partial \mu} (\ln L) &= 0 - \frac{1}{2\theta} \sum_{i=1}^{n} 2(X_i - \mu)(-1) \\ &= \frac{1}{\theta} \sum (X_i - \mu) \\ &= 0 \quad \text{when} \quad \sum X_i - n\mu = 0 \to \widehat{\mu} = \bar{X} \end{aligned}$$

$$\text{i.e. } \mu = \frac{1}{n}\sum X_i = \bar{X}$$

$$\begin{aligned}
\frac{\partial}{\partial\theta}(\ln L) &= -\frac{n}{2}\cdot\frac{2\pi}{2\pi\theta} + \frac{1}{2\theta^2}\sum_{i=1}^{n}(X_i-\mu)^2 \\
&= -\frac{n}{2\theta} + \frac{1}{2\theta^2}\sum_{i=1}^{n}(X_i-\mu)^2 \\
&= 0 \quad \text{At our potential critical point } \mu = \bar{X} \\
&\iff -n\theta + \sum_{i=1}^{n}(X_i-\bar{X})^2 = 0 \\
\text{or } \widehat{\theta} &= \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2
\end{aligned}$$

our old biased estimator for σ^2

Is this a max? We need to construct the Hessian Matrix...

$$\begin{aligned}
H &= \begin{bmatrix} \ln L_{\mu\mu} & \ln L_{\mu\theta} \\ \ln L_{\theta\mu} & \ln L_{\theta\theta} \end{bmatrix} \\
&= \begin{bmatrix} -\dfrac{n}{\widehat{\theta}} & -\dfrac{\sum X_i - n\mu}{\widehat{\theta}^2} \\ -\dfrac{\sum X_i - n\mu}{\widehat{\theta}^2} & \dfrac{n}{2\widehat{\theta}^2} - \dfrac{1}{\widehat{\theta}^3}\sum(X_i-\mu)^2 \end{bmatrix}
\end{aligned}$$

At $(\widehat{\mu},\ \widehat{\theta})$

$$H(\mu,\theta) = \begin{bmatrix} -\dfrac{n}{\widehat{\theta}} & 0 \\ 0 & \dfrac{n}{2\widehat{\theta}^2} - \dfrac{1}{\widehat{\theta}^3}\sum(X_i-\mu)^2 \end{bmatrix}$$

For a maximum, by the 2nd derivitive test, we need $(\ln L)_{\mu\mu} = -\dfrac{n}{\widehat{\theta}} < 0$, which it is! And $\det H > 0$.

Note, $(\ln L)_{\theta\theta} = \dfrac{\widehat{\theta}n - 2\sum(X_i-\bar{X})^2}{2\widehat{\theta}^3}$.

But $n\widehat{\theta} = \sum(X_i-\bar{X})^2$, so $(\ln L)_{\theta\theta} = \dfrac{\widehat{\theta}n - 2\widehat{\theta}n}{2\widehat{\theta}^3} = -\dfrac{n}{2\widehat{\theta}^2} < 0$.

So $\det H = \dfrac{-n}{\widehat{\theta}}\cdot\dfrac{-n}{2\widehat{\theta}^2} = \dfrac{n^2}{2\widehat{\theta}^3} > 0$. Hence $(\bar{X},\ \widehat{\theta})$ is the location of a MLE by the 2nd derivitive test.

Discussion: Why MLEs? Because they have nice properties.

① If U is a sufficient statistic of θ, then the MLE isa function of U.

Reason: If U is sufficient, then $L(\vec{x} \mid \theta)$ is factorable:

$$L(x_1,\ldots,x_n \mid \theta_1,\ldots,\theta_k) = g(U,\theta)\cdot h(x_1,\ldots,x_n)$$

This $\dfrac{\partial L}{\partial \theta} = \dfrac{\partial g}{\partial \theta} \cdot h(\vec{x})$

$$\text{i.e.} \quad \underbrace{\frac{\partial L}{\partial \theta} = 0}_{\substack{\text{sol'n to this is} \\ \widehat{\theta} \text{ (the MLE)}}} \iff \underbrace{\frac{\partial g}{\partial \theta} = 0}_{\substack{\text{sol'n to this is} \\ \widehat{\theta} \text{ a function of } U}}$$

② The invariance properties of the MLE.

Suppose we want to estimate a function of the parameter (say $t(\theta)$). If t is a one-to-one (injective) function (i.e. invertible), the MLE of $t(\theta)$ will be simply $t(\widehat{\theta})$, where $\widehat{\theta}$ is the MLE of θ.

Reason: Let $t(\theta)$ be injective. Assume $\widehat{\theta}$ the MLE of θ $t(\theta)$ invertible $\implies$ $\Psi = t(\widehat{\theta})$ and $t^{-1}(\Psi) = \theta$

Then $L(\vec{x} \mid \theta)$ is maximized at $\theta = \widehat{\theta}$. Then $L(\vec{x} \mid t^{-1}(\Psi))$ is maximized at the same $\widehat{\theta}$. Hence $t^{-1}(\Psi) = \widehat{\theta}$ or $\widehat{\Psi} = t(\widehat{\theta})$.

Example: (#1) Show $\bar{X}$ is MLE for λ when $X \sim \text{Pois}(\lambda)$.

In showing $S := \sum X_i$ is sufficient, we had factored the likelihood function

$$L(\vec{x} \mid \lambda) = \underbrace{e^{-n\lambda}\lambda^S}_{g(S,\,\lambda)} \cdot \frac{1}{x_1! \cdot x_2! \cdots x_n!}$$

To find the MLE $\dfrac{\partial L}{\partial \lambda} = 0$ when $\dfrac{\partial g}{\partial \lambda} = 0$.

$$\begin{aligned} \frac{\partial g}{\partial \lambda} &= -ne^{n\lambda}\lambda^S + e^{-n\lambda} s\lambda^{S-1} \\ &= e^{-n\lambda}\lambda^{S-1}\big[-n\lambda + S\big] \\ &= 0 \\ &\implies \lambda = \frac{S}{n} \qquad\qquad \text{i.e. } \widehat{\lambda} = \bar{x} \end{aligned}$$

As before, can show a max by the 1st derivitive test.

Example: Lots of bits together

We have shown that $\chi^2 \sim \text{Pois}(\lambda)$ with $S = \sum x_i$ is a sufficient statistic for λ. And we just showed that $\widehat{\theta} = \bar{x}$ is MLE for λ.

Hence, by the Rao-Blackwell Theorem, $\text{E}\left(\bar{X} \mid S\right) = \bar{X}$, then $\bar{X}$ is an MVUE for λ. At the end of RBT, via

$$w = \begin{cases} 1 & x_1 = 0 \\ 0 & x_1 \neq 0 \end{cases} \quad (w \text{ unbiased for } e^{-\lambda})$$

Combining w with S we get the MVUE $\widehat{\Psi} = \left(\frac{n-1}{n}\right)^S = \left(\frac{n-1}{n}\right)^{n\bar{X}}$ for $e^{-\lambda}$. (also $\text{Var}(\widehat{\Psi}) < \text{Var}(w)$) by RBT.

Now, since $\bar{X}$ is MLE for λ, by the invariance properties of MLE,

$$\theta^* = e^{-\bar{X}} \text{ is MLE for } e^{-\lambda}.$$

Reason: $g(x) = e^- x$ is clearly injective on $\mathbb{R}$.

Moreover, by RBT $\text{Var}(\widehat{\Psi}) < \text{Var}\left(e^{-\bar{X}}\right)$. And we have a conjecture that $e^{-\bar{X}}$ is actually biased.

While $e^{-\bar{X}}$ is the "natural" estimator for $e^{-\lambda}$, the RBT says $\widehat{\Psi} = \left(\frac{n-1}{n}\right)^{n\bar{X}}$ is "better" as $\widehat{\Psi}$ is unbiased and $\text{Var}(\widehat{\Psi}) < \text{Var}(\theta^*)$.

Q: Can we show this directly?
Maybe. If θ^* is unbiased...

$$\text{Var}(\widehat{\Psi}) < \text{Var}(\theta^*) \leftrightsquigarrow \text{E}\left[(\widehat{\Psi})^2\right] < \text{E}\left[(\theta^*)^2\right]$$

FACT: $\lim_{n\to\infty}\left(\frac{n-1}{n}\right)^n = \lim_{n\to\infty}\left(1-\frac{1}{n}\right)^n = e^{-1}$. (This is via Calc I application of L'Hopitals)

Moreover, via the same ln technique, it can be shown that $f(x) = \left(\frac{x-1}{x}\right)^x$ is increasing on $x \geq 2$ (see addendum). Hence,

$$\begin{aligned}
&\left(\frac{n-1}{n}\right)^n < e^{-1} \\
\implies &\left(\frac{n-1}{n}\right)^{2n} < e^{-2} \\
\implies &\left(\frac{n-1}{n}\right)^{2n\bar{X}} < e^{-2\bar{X}} \\
\implies &\left(\left(\frac{n-1}{n}\right)^{n\bar{X}}\right)^2 < e^{-\bar{X}^2} \\
\implies &\text{E}\left(\widehat{\Psi}^2\right) < \text{E}\left((\theta^*)^2\right) \\
\implies &\text{Var}\left(\widehat{\Psi}\right) < \text{Var}\left(\theta^*\right) \\
&\text{(if } \theta^* \text{ is unbiased)}
\end{aligned}$$

Addendum:

Claim: $f(x) = \left(\dfrac{x-1}{x}\right)^x$ is increasing on $x \geq 2$.

Reason:

$$\begin{aligned} \ln f &= x\ln(x-1) - x\ln x \\ \frac{f'}{f} &= \ln(x-1) + \frac{x}{x-1} - \ln x - 1 \\ &= \ln\left(\frac{x-1}{x}\right) + \frac{1}{x-1} \\ f'(x) = 0 \implies \ln\left(\frac{x-1}{x}\right) &+ \frac{1}{x-1} = 0 \end{aligned}$$

But this does not happen on $(1,\infty)$. Equivalently, $\ln\left(\dfrac{x-1}{x}\right)^{1-x} = 1 \implies \dfrac{x-1}{x} = 1$ has no solutions.

Since $f'(2) = \ln\left(\dfrac{1}{2}\right) + 1 = 1 - \ln 2 > 0$, we get $f'(x) > 0$ on $(1,\infty)$.

Chapter 10

Hypothesis Testing

Motivational example:

I claim that I am a 75% free-throw shooter. You need to verify, and make me shoot 20 free throws. I make 8.

You are suspicious of my claim.

Is that suspicion founded?

We can think of a free-throw as Bern(0.75).

The probability of getting 8 is binomial.

$$\Pr(X = 8) = \binom{20}{8}\left(\frac{3}{4}\right)^{8}\left(\frac{1}{4}\right)^{12} = 0.00075$$

This implies that if $p = \frac{3}{4}$ is true, we observed a "rare event."

But, another way to think about this is if $p = \frac{3}{4}$, what is the probability that I would shoot 8 or worse?

$$\Pr(X \leq 8) = \sum_{i=0}^{8}\binom{20}{i}\left(\frac{3}{4}\right)^{i}\left(\frac{1}{4}\right)^{20-i} < 0.001$$

The entire array, $0, 1, \ldots, 8$ is less than $\frac{1}{1000}$ probability. We are "safe" to conclude that the true free-throw percentage is less than 75%.

This is what hypothesis testing is all about.

- Assume something is true
- Observe and compute the probabilty of this outcome, or more extreme (rarer) events.
- Decide if there is evidence against your assumption.

Flip-side: In chapter 8 we would have constructed a confidence interval for the true p.

$n = 20$ is small... we need $t(\text{df} = 19)$ distribution.

A 99% C.I. for p

Uses $t_{0.005}(19) = 2.861$ standard errors.

$$\widehat{p} \pm t_{0.005}\sqrt{\frac{\widehat{p}\widehat{q}}{20}} \approx 0.4 \pm 0.256 \implies p \in (0.144, 0.656)$$

$p = \frac{3}{4}$ is very far outside. Again, very unlikely that $p = \frac{3}{4}$.
*C.I. and the Hypothesis Test are 2-sides of the same coin.

Hypothesis Testing Vocab

① The Hypotheses.

- **The null hypothesis** H_0 : Assumption of no change.
  Assumed true unless "proven" otherwise.
  Example: $H_0 : p_0 = 0.75$

- **The alternative hypothesis** H_a : The claim we aim to test.
  $H_a : p < 0.75$
  Example (One sided): $H_a : p < p_0$ or $H_a : p > p_0$

  Example (Two sided): $H_a : p \neq p_0$

② **The decision rule**. This is the cutoff α we use to accept or reject H_a. Usually stated beforehand.

③ **The decision**. After a probability calculation, either

(i) Reject the null hypothesis H_0 or

(ii) Not enough evidence to reject H_0 (accept H_a)

④ ***p*-value.** The probability of seeing as much, or more evidence for H_a than we saw in the data.

$$p = \Pr(X \leq 8) < 0.001$$

*The smaller the p-value, the more support for H_a.
Test Results are called **statistically significant** if H_0 is rejected.

Topic: Error Types

2 possible decisions $\implies$ 2 possible mistakes.

The "truth"	H_0	H_a
Test Supports H_0	good	Type II
Test Supports H_a	Type I	good

Type I: We reject H_0 when H_0 is the truth.
*This error comes from the decision rule.
Example: I actually am a 75% free-throw shooter, but had a bad day.

Type II: We accept H_0 when H_a is the truth.

We will see that we can measure these error types in some sense, but it depends on our decision rule α and the unknown, true population parameter θ.

$$\Pr(\text{Type I error}) = \alpha.$$
$$\Pr(\text{Type II error}) = \beta.$$

Example: Back to the free throws.

Decision Rule: Beforehand, you decide if I make 10 or less, you will reject $H_0 : p = 0.75$.

The text calls the set of outcomes the **rejection region** (RR). If $T \in \text{RR}$ then we reject H_0 (support the alternative).

Here, $RR = \{0, 1, 2, \ldots, 10\}$ and we can exactly calculate this:

$$\begin{aligned}
\alpha &= \Pr(\text{Type I error}) \\
&= \Pr(\text{rejecting } H_0 \text{ when } H_0 \text{ is true}) \\
&= \Pr\left(0 \leq T \leq 10 \text{ and } p = \frac{3}{4}\right) \\
&= \text{pbinom}(10,\ 20,\ 0.75) \\
&= 0.013
\end{aligned}$$

Hence, it is very unlikely to get a Type I error.

Computing Type II error probabilities requires a guess for the true H_a.

$$\begin{aligned}
\beta &= \Pr(\text{Type II error}) \\
&= \Pr(\text{accept } H_0 \text{ when } H_a \text{ is true}) \\
&= \Pr\left(11 < T \leq 20 \text{ when } p = \theta < \frac{3}{4}\right) \\
&= 1 - \Pr\left(0 \leq T \leq 10 \text{ when } p = \theta < \frac{3}{4}\right) \\
\beta(\theta) &= 1 - \text{pbinom}(10,\ 20,\ \theta) \quad \text{a function of } \theta
\end{aligned}$$

Here, if $\theta = 0.6$, then $\underbrace{\beta = 0.755}_{\text{This is a lot, 75\%}}$

If $\theta = 0.5$, $\beta = 0.4119$

Note the larger the true difference between θ and $p_0 = \frac{3}{4}$ is, the smaller the Type II error.

FACT: α and β are inversely related! If we increase α (Probability of Type I error), we see a decrease in β (Probability of Type II error) (and vice versa).

Definition: The **power function** of a test is defined as

$$\text{power}(\theta) = 1 - \beta(\theta)$$

and this measures Type II error.

Example: New ultrasound machine: Claimed to detect dumors better than the old machine.

Hospital designs a test: Take a known patient with a known tumor distribution "tumor set." Scan each patient in both machine: record the proportion of known tumors detected with each machine.

Let p_0 be the proportion found of old machine, and p_1 be the new.

(a) $H_0 : p_0 = p_1$ and $H_a : p_0 < p_1$

(b) A Type I error occurs when we decide the new machine is better, when it is actually worse.

Real world consequences: Results in investment in "better" equipment when is not better and does not help your patients

(c) A Type II error occurs when we accept H_0 when H_a is true. (We think that the old equipment is better when it isn't.)

Real world consequences: We could have had better machines, detected more cancer, and saved more lives (but didn't).

10.3 Z-tests (large samples)

Example (10.18): The hourly wages in a particular industry is distrubuted $N(13.20,\ 2.50)$. A company in this industry emplpoys 40 workers, paying them an average of \$12.20 per hour. Can this company be accused of paying in substandard wages. Use $\alpha = 0.01$.

Solution: Recall $n = 40 > 30$ is considered "large", Hence

$$\begin{aligned}\bar{X} &\sim N\left(\mu, \frac{\sigma^2}{n}\right) \\ &= N\left(13.20, \frac{2.50}{40}\right) \\ &= N\left(13.20, 0.0625\right)\end{aligned}$$

Our test:

$$H_0 : \mu_c = 13.20 \qquad\qquad H_a : \mu_c < 13.20$$

Where μ_c is the company average. Here, $\alpha = 0.01$ is the decision rule, also the significance level, and also the probability of making a Type I error (reject H_0 when H_0 is true).

$$\begin{aligned}\alpha &= \Pr(\text{Type I error}) \\ &= \Pr(\text{rejecting } H_0 \text{ when it is true}) \\ &= 0.01\end{aligned}$$

Compute the p-value.

$$p = \Pr\left(\bar{X} \leq 12.20 \mid \mu = 13.20\right)$$

Convert to a Z-score (like we did in 325).

$$= \Pr\left(Z \leq \frac{12.20 - 13.20}{0.25}\right)$$

$$Z = \frac{\bar{X} - \mu}{\sigma - \bar{X}} \qquad\qquad \sigma_{\bar{X}}\sqrt{\frac{2.50}{40}} = \frac{1}{4}$$

$$\begin{aligned}p &= \Pr(Z \leq -4) \\ &= \Pr(Z \geq 4) \\ &= 0.0000317 \text{ by table 4} \\ &= \text{pnorm}(12.20, 13.20, 0.25)\end{aligned}$$

Decision and conclusion: Since p-value is much less than $0.01 = \alpha$, we reject H_0 (accept the alternative). Therefore we conclude that, yes, the company appears to be systematically underpaying its employees in relation to the rest of the industry.

Remark:

① $Z = \dfrac{\bar{X} - \mu}{\sigma - \bar{X}} = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ is called the test statistic or Z-statistic. E.g. $Z = -4$.

② R command: $\text{pnorm}(0, \mu, \sigma)$ and is a left tail calculator (opposite of Table 4).

Example (10.21): Shear strength measurements are derived from unconfined compression tests for two types of soils.

Soil 1	**Soil 2**
$n_1 = 30$	$n_2 = 35$
$\bar{Y}_1 = 1.65$	$\bar{Y}_2 = 1.43$
$S_1 = 0.26$	$S_2 = 0.22$

Tons per square foot (unit).

Do the soils appear to differ with respect to average shear strength at the 1% significance level ($\alpha = 0.01$).

Note that $n_1, n_2 \geq 30$. This implies that we can use "large sample" assumptions.

i.e. $\sigma_1 = S_1$ and $\sigma_2 = S_2$ without loss of precision (i.e. no need for t-distribution).

$$H_0 : \mu_1 = \mu_2 \implies \mu_1 - \mu_2 = 0$$

$$H_a : \mu_1 \neq \mu_2 \quad \leftarrow \text{This is called a 2-sided test}$$

In Chapter 8, we saw $\bar{Y}_1 - \bar{Y}_2$ is distrubuted

$$N\left(\mu_1 - \mu_2\ ,\ \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

Under null hypothesis,

$$N\left(0\ ,\ \frac{0.26^2}{30} + \frac{0.22^2}{35}\right)$$
$$= N(0, 0.00363)$$

and $\sigma_{\bar{Y}_1 - \bar{Y}_2} = 0.0603$

Note that $\bar{Y}_1 - \bar{Y}_2 = 1.65 - 1.43 = 0.22$

$$\begin{aligned}
p &= \Pr\left(|\bar{Y}_1 - \bar{Y}_2 - 0| > 0.22\right) \\
&= \Pr\left(\bar{Y}_1 - \bar{Y}_2 < -0.22\right) + \Pr\left(\bar{Y}_1 - \bar{Y}_2 > 0.22\right) \\
&= 2\Pr\left(\bar{Y}_1 - \bar{Y}_2 > 0.22\right) \\
&= 2\Pr\left(Z > \frac{0.22}{0.0603}\right) \quad \text{not on table 4} \\
&< 2\Pr\left(Z > 3.5\right) \\
&= 2 \cdot 0.000233 \\
&= 0.000466 \\
&< \alpha
\end{aligned}$$

Conclusion: This is statistically significant. I.e. supports H_a. The sheer strengths are different.

Remark:

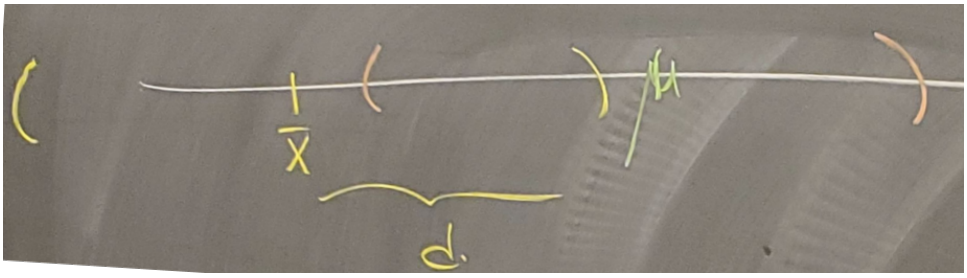
① $p = 2\,\text{pnorm}(-0.22, 0, 0.0603) = 0.000265$

② In the last 2 examples, using the emperical rule (68-95-99.7), we could have concluded "Reject H_0" simply on the Z-score alone. (Once we're more than 3σ away from μ, $p < 100\% - 99.7\% = 0.3\%$).

$$\Pr\left(Z < Z_*\right) = \alpha \qquad \text{and} \qquad H_0 : \mu' = \mu \qquad \text{and} \qquad H_a : \mu' < \mu.$$
$$-Z_* = \frac{C^* - \mu}{\sigma/n} \quad \iff \quad C^* = \mu - Z_* \frac{\sigma}{\sqrt{n}}$$

The rejection region RR is $\bar{X} \leq C^*$

$\mu_0 \pm Z^* \dfrac{\sigma}{\sqrt{n}}$ is a $1 - \alpha$ level C.I. If $\bar{X}$ lands in this interval, it supports H_0, else if it lands outside, reject H_0.



10.4 More about errors and sample size

Motivational example: X equals the breaking strength of a steel bar. If the bar is manufactured by Process I, it is known $X \sim N(50,\ 36)$.

We now have Process II and it is hoped that the steel is 10% stronger. i.e. $X \sim N(55,\ 36)$.

Our test?

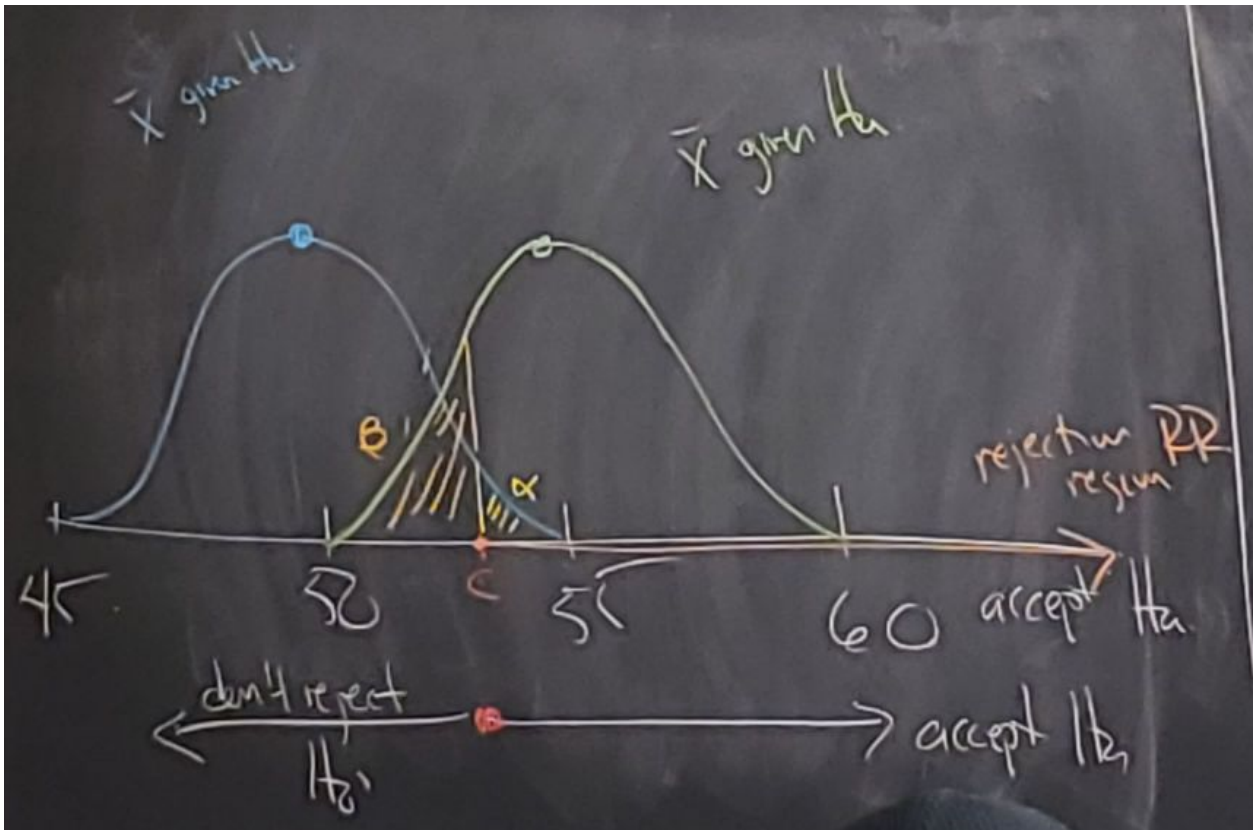
$$H_0 : \mu_I = 50 \qquad\qquad H_a : \mu_{II} = 55$$

Okay... we can't really test "this"

But we can construct a hypothetical test where if H_a is true, we can minimize (or control) both the Type I and Type II errors.

For the sake of concrete-ness, set $n = 16$. Then,

$$\sigma^2_{\bar{X}} = \frac{36}{16} \qquad \text{and} \qquad \sigma_{\bar{X}} = 1.5$$



$$\alpha = \Pr(\text{Type I}) \qquad H_0 : \mu = 50 \qquad H_a : \mu < 50$$

On the other hand, given α, we can also see

$$\beta = \Pr(\text{Type II}) = \Pr(\text{accept } H_0 \text{ when } H_a \text{ is true})$$

Given α, we can find c.

$$\begin{aligned}\alpha &= \Pr\left(\bar{X} > C \mid H_0\right) \\ &= \Pr\left(\frac{\bar{X} - 50}{1.5} > \frac{C - 50}{1.5}\right)\end{aligned}$$

and

$$\begin{aligned}\beta &= \Pr\left(\bar{X} < C \mid H_a\right) \\ &= \Pr\left(\frac{\bar{X} - 55}{1.5} < \frac{C - 55}{1.5}\right)\end{aligned}$$

For fixed $n = 16$, usually choose α small.

$$\alpha = \Pr\left(\bar{X} - 50 > 2\sigma_{\bar{X}}\right) = 0.025$$

Then $C = 50 + 2(1.5) = 53$ and

$$\begin{aligned}\beta &= \Pr\left(\bar{X} < 53 \mid H_a\right) \\ &= \Pr\left(\frac{\bar{X} - 55}{1.5} < 1.33\right) \\ &= 0.0918\end{aligned}$$

Note almost four times as likely to make a Type II error than a Type I error. Of course, decreasing α with increase β.

$$\alpha = 0.01 \quad \Longrightarrow \quad Z\text{-score} = 2.33 \qquad (Z_{0.98})$$

$$c = 50 + 2.33(1.5) = 53.495$$

$$\begin{aligned}\beta &= \Pr\left(\bar{X} < 53.495 \mid H_a\right) \\ &= \Pr\left(Z < -1.003\right) \\ &= 0.1587\end{aligned}$$

Again, we note that the only way to decrease *both* α and β is to crank up the n.

Discussion: Choosing sample size n. We consider the 1-sided test

$$H_0 : \mu = \mu_0 \qquad\qquad H_a : \mu > \mu_0$$

Fix α at the start.

$$\begin{aligned}\alpha &= \Pr\left(\bar{X} > C \text{ when } \mu = \mu_0\right) \\ &= \Pr\left(Z > \frac{C - \mu_0}{\sigma/\sqrt{n}}\right) \\ &= \Pr\left(Z > Z_\alpha\right) \qquad Z_\alpha = \frac{C - \mu_0}{\sigma/\sqrt{n}}\end{aligned}$$

But, as with the power function, we need to choose specific μ_a's to work with (e.g. $\mu_a = 55$).

$$\begin{aligned}\beta &= \Pr\left(\bar{X} < C \text{ when } \mu = \mu_a\right) \\ &= \Pr\left(Z < \frac{C - \mu_a}{\sigma/\sqrt{n}}\right) \\ &= \Pr\left(Z < -Z_\beta\right) \quad \text{when} \quad -Z_\beta = \frac{C - \mu_\alpha}{\sigma/\sqrt{n}}\end{aligned}$$

2 equations in the 2 unknows, $C,\ n$

$$C = \mu_0 + Z_\alpha \frac{\sigma}{\sqrt{n}} \qquad \text{and} \qquad C = \mu_a - Z_\beta \frac{\sigma}{\sqrt{n}}$$

$$\mu + Z_\alpha \frac{\sigma}{\sqrt{n}} = \mu_a - Z_\beta \frac{\sigma}{\sqrt{n}}$$

Solve for n:

$$(Z_\alpha + Z_\beta)\frac{\sigma}{\sqrt{n}} = \mu_a - \mu_0$$

$$n = \frac{(Z_\alpha + Z_\beta)^2 \sigma^2}{(\mu_a - \mu_0)^2}$$

Remark:

① Of course all of this the fudge factor that we don't really know μ_a.

② If we did $H_0 : \mu_0 = \mu_a$ and $H_a : \mu_0 > \mu_a$ we get the same formula for n. "sample size estimator for a one-sided α-level test"

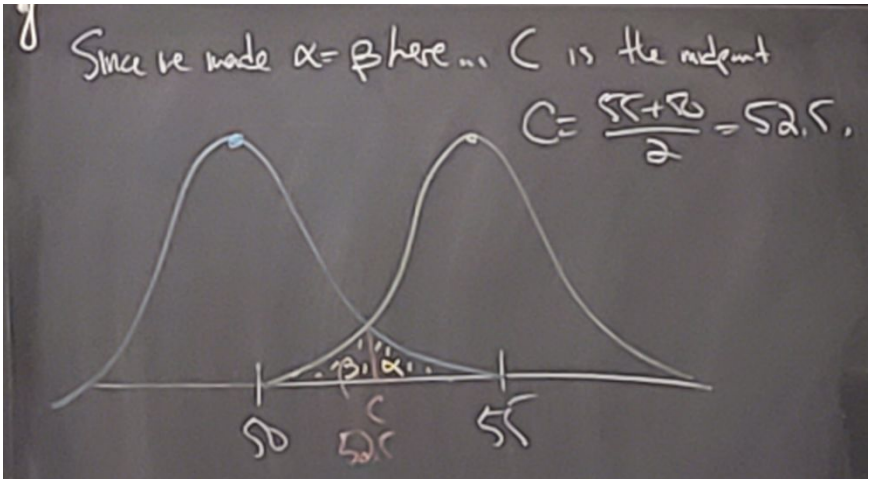
Example: Back to the steel example

If we decided at the start that we want $\alpha = \beta = 0.05$, what n should we choose?

For $\alpha = 0.05 \implies Z_\alpha = Z_{0.05} = 1.645$. Similarly for β, we need $Z_\beta = 1.645$. Into our formula,

$$n = \left(\frac{1.645 + 1.645}{55 - 50}\right)^2 \cdot 36 = \lceil 15.5867 \rceil = 16$$

Since we made $\alpha = \beta$ here, C is the midpoint $C = \dfrac{55 + 50}{2} = 52.5$.



10.8 T tests

Recall that for a small sample $n < 30$. We need to use the t-distribution.

Chapter 8: t-stat confidence interval $\bar{X} \pm t_{\alpha/2}(\text{df})\sqrt{\frac{s^2}{n}}$

Example: 100 mL sample of water from swimming areas are tested for fecal colliform bacteria. It is considered to be safe if the level of bacteria is less than 400 in 3.3 oz.

20 Samples are taken: Found $\bar{X} = 1231$ and $S = 1038$.

Construct a test to determine if it's safe to swim. Our test:

$$H_0 : \mu_0 = 400$$
$$H_a : \mu_a > 400$$

Test statistic $T = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}} = \dfrac{1231 - 1038}{1038/\sqrt{20}} = 0.350$ is the test statistic.

The degrees of freedom is $n - 1 = 19$. Using table 5, $t_{0.005}(19) = 2.861$ The p-value is:

$$\begin{aligned} p &= \Pr(T \geq 3.580 \mid \mu_0 = 400) \\ &< \Pr(T \geq 2.861) \\ &= 0.005 \end{aligned}$$

Reject H_0: There is poop in the water; stay out!

Remark: In general:

$$H_a := \begin{cases} \mu > \mu_0 \\ \mu < \mu_0 \\ \mu \neq \mu_0 \end{cases} \implies \text{RR} := \underbrace{\begin{cases} t > t_\alpha \\ t < -t_\alpha \\ |t| > t_{\alpha/2} \end{cases}}_{\text{the } t\text{-stat}}$$

Example: Lifestyle comparison. Monitoring the active time (in minutes per day) between 2 populations: obese and lean.

Group	**Count**	**Standing/Walking**	S
Obese	$n = 10$	373.269	67.498
Lean	$n = 13$	525.751	107.121

Question: Are these two groups significant significant? This is a "are population means different" question.

$$\begin{aligned} H_0 : \mu_L = \mu_0 \quad &\implies \quad \mu_L - \mu_0 = 0 \\ H_a : \mu_L \neq \mu_0 \quad &\implies \quad \mu_L - \mu_0 \neq 0 \end{aligned}$$

Questions about "different" results in a 2-sided test

Confidence interval: $\bar{X}_1 - \bar{X}_2 \pm t_{\alpha/2}(\text{df}) \left(S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right)$

$$S_{\text{pool}} = \sqrt{\frac{(n_1 - 1){S_1}^2 + (n_2 - 1){S_2}^2}{n_1 + n_2 - 2}} \quad \text{with df} = n_1 + n_2 - 2$$

The difference in means t test statistic is

$$T = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Our $\bar{X}_L - \bar{X}_0 = 152.482$. $H_0 : \mu_L - \mu_0 = 0$ is the null hypothesis. $\text{df} = 10 + 13 - 2 = 21$.

$$\begin{aligned} S_p &= \sqrt{\frac{(13-1)107.121^2 + (10-1)67.498^2}{13 + 10 - 2}} \\ &= 92.247. \end{aligned}$$

$$\begin{aligned} T &= \frac{152.481}{92.247 \sqrt{\frac{1}{10} + \frac{1}{13}}} \\ &= 3.93 \end{aligned}$$

$$\begin{aligned} p &= \Pr(|T| > 3.93) \\ &= 2 \cdot \Pr(T > 3.93) \\ &< 2 \cdot 0.0005 \qquad \text{(by table 5)} \\ &= 0.001 \end{aligned}$$

So p is super small ($p < \alpha$) which implies that we reject the null hypothesis. That is, $\mu_0 < \mu_L$ is true.

Our goal this week is to justify hypothesis testing.

Last day: Pearson Neyman Lamma

$X_1, \ldots, X_n \sim$ via PDF $f(x \mid \theta)$ when θ_0 and θ_a are two possible values of θ. If there exists a the constant k and a subset C of the sample space such that

1. $P\{(x_1, \ldots, x_n) \in C \mid \theta_0\} = \alpha$
2. $\frac{L(\theta_0)}{L(\theta_a)} \leq k$ for $x_1, \ldots, x_n \in C$.
3. $\frac{L(\theta_0)}{L(\theta_a)} \geq k$ for $x_1, \ldots, x_n \in \overline{C}$

Then C is a best critical region of size α for testing $H_0 : \theta = \theta_0$ versus $H_a : \theta = \theta_a$.

Before the proof... example.

Example: Let $Y_1, \ldots, Y_n$ from $f(y \mid \theta) = \dfrac{2}{\theta} y e^{-y^2/\theta}$ and $y > 0$. The Rayleight distribution.

Want to test $H_0 : \theta = \theta_0$ versus $H_a : \theta = \theta_a$.

Note $L(Y_1, \ldots, Y_n \mid \theta) = \dfrac{2}{\theta} y_i e^{-y_i^2/\theta} \cdots \dfrac{2}{\theta} y_n e^{-y_n^2/\theta}$.

$$\left(\frac{2}{\theta}\right)^n \underbrace{\exp\left[-\left(\sum_1^n y_i^2\right) \Big/ \theta\right]}_{g(S,\theta)} \underbrace{\prod_1^n y_i}_{h}$$

We will come back to this and connect to sufficiency next day.

The N-P ratio of likelihood function becomes

$$\begin{aligned}\frac{L(\theta_0)}{L(\theta_a)} &= \frac{\left(\frac{2}{\theta_0}\right)^n \exp\left[-\left(\sum_1^n y_i^2\right) \Big/ \theta_0\right] \prod_1^n y_i}{\left(\frac{2}{\theta_a}\right)^n \exp\left[-\left(\sum_1^n y_i^2\right) \Big/ \theta_a\right] \prod_1^n y_i} \\ &= \left(\frac{\theta_a}{\theta_0}\right)^n \exp\left[-\sum y_i^2 \cdot \left(\frac{1}{\theta_0} - \frac{1}{\theta_a}\right)\right]\end{aligned}$$

We want C to relate to our rejection region RR.

So (2), implies "reject H_0 if"

$$\left(\frac{\theta_a}{\theta_0}\right)^n \exp\left[-\sum y_i^2 \cdot \left(\frac{1}{\theta_0} - \frac{1}{\theta_a}\right)\right] \leq k$$

This looks scary, but θ_a and θ_0 are fixed and $\theta_0 < \theta_a$ so $\left(\frac{1}{\theta_0} - \frac{1}{\theta_a}\right) > 0$. In the end, making $\dfrac{L(\theta_0)}{L(\theta_a)}$ "small enough" for k simlifies does to the condition that $\sum_i^n y_i^2$ is large enough. i.e. we need $\sum_1^n y_i^2 > k'$ for some k' (dependent upon k).

New problem: To determine an appropriate k', we need to know how $S = \sum_1^n y_i^2$ is distributed.

Using CDF method (§ 6.3) we can show that the distribution of y^2 is exponention with mean θ. Then, via products of moment generating functions.

$$S = \sum_1^n y_i^2 \sim \text{Gamma}\left(n, \frac{1}{\theta}\right)$$

Stop! What just happened.

The N-P lemma tells us that the most powerful test of H_0 vs H_a is to use the statistic $S = \sum y_i^2$. Then, given any α level significance (5%, 1%, whatever), we use the test on S to be if S is larger than $100(1-\alpha)$ percentile of the Gamma $\left(n, \frac{1}{\theta}\right)$ distribution... we reject H_0.

Picture with some γ_* such that $\int_{\gamma_*}^{\infty} = \alpha$.

$S > \gamma_*$, we reject the null hypothesis. Moreover, by the NP lemme, we don't have to compare the powe rof other possible tests because the lemma says any data

$$(Y_1, \ldots, Y_n) \in C \leftrightsquigarrow S \geq \gamma_*$$

$$C := \left\{ (Y_1, \ldots, Y_n) \mid \sum y_i^2 \geq \gamma_* \right\}$$

which is itself a consequence of our significance level α results in the most powerful test.

$$P(C \mid \theta_a) \geq P(D \mid \theta_a)$$

where D is another critical region ($P(D \mid \theta_a) = \alpha$).

Definition: The test using the best critical region is called the most powerful test.

Recap: We wanted to construct a hypothesis test at α significance level. All in one fell swoop, the N-P lemme says

1. Here is the stat to use ($S = \sum y_i^2$ in last example)
2. Here is your rejection region ($S \geq \gamma_a$ in last example)
3. in fact, this is the most powerful test possible.

(This is kinda awesome)
Theorem: NP lemma;

Proof. (Continuous)
Let $B \subseteq \mathbb{R}^n$ and define $\underbrace{\int_B L(\theta) = \int \cdots \int}_{n\text{times}} L(x_1, \ldots, x_n \mid \theta)\, dx_1\, dx_2, \ldots, dx_n$.
Assume there exists a critical region C satisfying bullets 1 2 and 3.

α fixed, $\alpha = \int_C L(\theta_0)$ (same as $P(C \mid \theta_0)$)

Need to prove "best", i.e. $P(C \mid \theta_a) \geq P(D \mid \theta_a)$

Assume D is another critical region,

$$\alpha = \int_D L(\theta)_0 = P(D \mid \theta_0)$$

So $0 = \int_C L(\theta_0) - \int_D L(\theta_0)$

$$
\begin{aligned}
0 &= \int_{C\cap D^C} L(\theta_0) + \int_{C\cap D} L(\theta_0) - \left[\int_{C\cap D} L(\theta_0) + \int_{C^C\cap D} L(\theta_0)\right] \\
&= \int_{C\cap D^C} L(\theta_0) - \int_{C^C\cap D} L(\theta_0)
\end{aligned}
$$

By (2), there exists k such that $\dfrac{L(\theta_0)}{L(\theta_a)} \le k$ i.e. $L(\theta_0) \le kL(\theta_a)$ at every point in C.

$$
\begin{aligned}
&\le k\int_{C\cap D^C} L(\theta_a) - \int_{C^C\cap D} L(\theta_0)
\end{aligned}
$$

By (3) there exists k such that $\dfrac{L(\theta_0)}{L(\theta_a)} \ge k$ i.e. $L(\theta_0) \ge kL(\theta_a)$

$$
\begin{aligned}
&\le k\left[\int_{C\cap D^C} L(\theta_a) - \int_{C^C\cap D} L(\theta_a)\right] \\
&= k\left[\int_{C\cap D^C} L(\theta_a) + \int_{C\cap D} L(\theta_a) - \left[\int_{C\cap D} L(\theta_a) + \int_{C^C\cap D} L(\theta_a)\right]\right] \\
&= k\left[\int_C L(\theta_a) - \int_D L(\theta_a)\right] \\
&\implies \int_C L(\theta_a) - \int_D L(\theta_a) \ge 0 \\
&\implies \int_C L(\theta_a) \ge \int_D L(\theta_a) \\
&\implies P(C \mid \theta_a) \ge P(D \mid \theta_a)
\end{aligned}
$$

Here, C is a best critical region of size α by defininition. □

Example: Back to our old sample size example.

$$X_1, \ldots, X_n \sim \mathrm{N}(\mu, 36)$$

We played with $H_0 : \mu = 50$ and $H_a : \mu = 55$.

Consider the ratio of the likelihood functions.

$$
\begin{aligned}
\frac{L(50)}{L(55)} &= \frac{(72\pi)^{-n/2}\exp\left[-\left(\frac{1}{72}\right)\sum_1^n (X_i-50)^2\right]}{(72\pi)^{-n/2}\exp\left[-\left(\frac{1}{72}\right)\sum_1^n (X_i-55)^2\right]} \\
&= \exp\left[-\frac{1}{72}\sum_1^n \left[(X_i-50)^2-(X_i-55)^2\right]\right] \\
&= \exp\left[-\frac{1}{72}\sum_1^n (10X_i-525)\right] \\
&= \exp\left[-\frac{5}{36}\sum_1^n X_i + \frac{175n}{24}\right] \leq k \text{ to satisfy (2)} \\
&\implies -\frac{5}{36}\sum_1^n X_i + \frac{175n}{24} \leq \ln k \\
&\implies \sum_1^n X_i \geq \frac{105n}{2} - \frac{36}{5}\ln k \\
&\implies \bar{X} = \frac{1}{n}\sum_1^n X_i \geq \frac{105}{2} - \frac{36\ln k}{5n}
\end{aligned}
$$

We have our stat $\bar{X} \geq k'$ thus will define our rejectoin rejection.

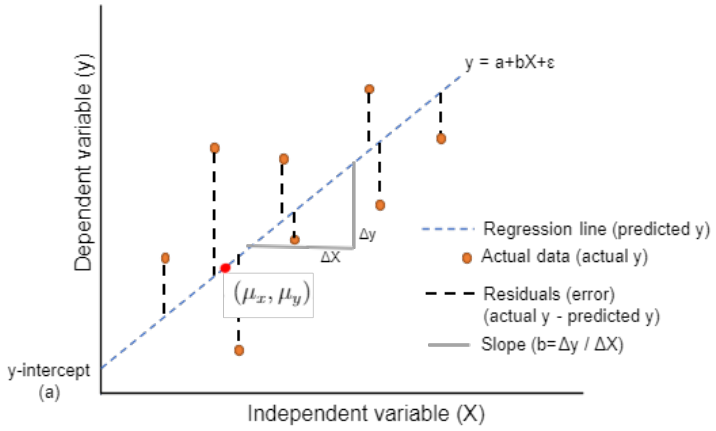
Chapter 11

Linear Models and Estimation by Least Squares

11.1 Introduction

We are going to do best-fit lines in 2 separate ways.

Consider some sample (data) $S = \{(x_i, y_i)\}$. In the background, there can be an underlying joint PDF $f(x, y)$



11.2 Linear Statistical Models

Assumptions

① Y is dependent upon X. (i.e. not an independent random variable.)

② In general, $y = mx + b$, the general relationship between X and Y.

11.3 The Method of Least Squares

A.) The probabalistic construction:

Given all possible lines $y = mx + b$ that *could* describe the data, the "most likely" one will be a line that intersects the point (μ_x, μ_y).

That is $y - \mu_y = m(x - \mu_x)$. Because we are interpreting y and a function of x, the "best line" should be one that minimizes the vertical distance (error (residuals)) in the y-direction.

For a point $(x_k,\, y_k)$ in S, the vertical distaince is

$$\text{dist} = \big|y_k - \big(m(x_k - \mu_x) + \mu_y\big)\big|.$$

Minimizing absolute values is a pain, but calculus I suggests we square this,

$$\text{dist}^2 = \Big((y_k - \mu_y) - m(x_k - \mu_x)\Big)^2.$$

As x, y are distributed via some PDF, the best way to compute our minimization problem is to minimize expectation with respect to slope m:

$$K(m) := \mathrm{E}\left(\text{dist}^2(m)\right) = \underbrace{\mathrm{E}\left(\Big((y_k - \mu_y) - m(x_k - \mu_x)\Big)^2\right)}_{**}.$$

** The m that minimizes is called the solution to our least-squares problem.

$$\begin{aligned} K(m) &= \mathrm{E}\left((y_k - \mu_y)^2 - 2m(x_k - \mu_x)(y_k - \mu_y) + m^2(x_k - \mu_x)^2\right) \\ &= \mathrm{E}\left((y_k - \mu_y)^2\right) - 2m\,\mathrm{E}\left((x_k - \mu_x)(y_k - \mu_y)\right) + m^2\,\mathrm{E}\left((x_k - \mu_x)^2\right) \\ &= S_y^2 - 2m\,\mathrm{Cov}(X, Y) + m^2 S_x^2 \end{aligned}$$

Minimizing (with respect to m) $\dfrac{\partial}{\partial m}$,

$$K'(m) = -2\,\mathrm{Cov}(X, Y) - 2mS_x^2.$$

With $K'(m) = 0$ then $m = \dfrac{\mathrm{Cov}(X, Y)}{S_x^2}$.

Note that by the second derivitive test, $K''(m) = 2S_x^2 > 0$. In AP Stats (MTH 111), the slope is moved to a "prettier" form:

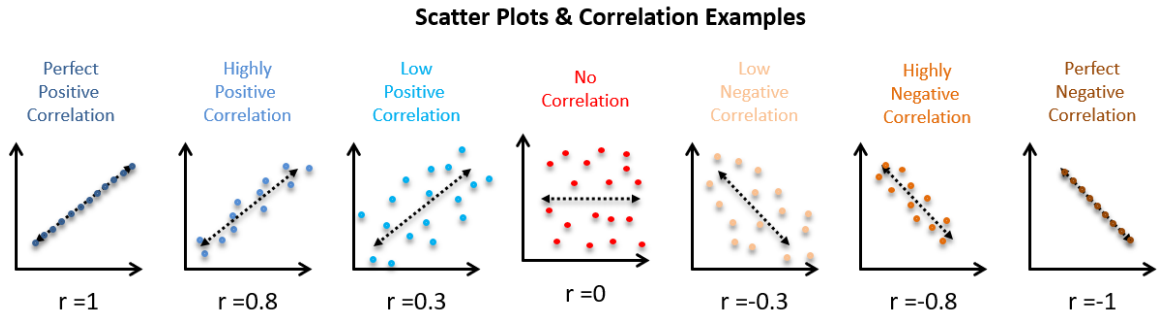
$$\begin{aligned} m &= \frac{\mathrm{Cov}(X, Y)}{S_x^2} \\ &= \frac{\mathrm{Cov}(X, Y)}{S_x S_y} \cdot \frac{S_y}{S_x} \\ &= \rho \cdot \frac{S_y}{S_x} \qquad\qquad \rho := \frac{\mathrm{Cov}(X, Y)}{S_x S_y} \end{aligned}$$

FACT #1 : The sign of the slope depends entirely upon ρ, where the sign is dependent upon $\text{Cov}(X, Y)$

FACT #2 : $-1 \leq \rho \leq 1$.

(For my MTH 325 class, we proved this fact in full generality).

FACT #3 : The closer $|\rho|$ is to 1, the better the line $\widehat{y} = mx + b$ "fits" the data.



The AP Stats definition fo the best fit line is $\widehat{y} = \widehat{b}_0 + \widehat{b}_1 x$ where $\widehat{b}_1 = \rho \dfrac{S_y}{S_x}$ and $\widehat{b}_0 = \bar{y} - \widehat{b}_1 \bar{x}$.

B.) The linear algebra construction:

We still have $S = \{(x_k,\, y_k)\}$. We want $y = mx + b$. Using S always yields an overdetermined (i.e. inconsistent) system

$$\begin{aligned} y_1 &= mx_1 + b \\ y_2 &= mx_2 + b \\ &\vdots \\ y_n &= mx_n + b. \end{aligned}$$

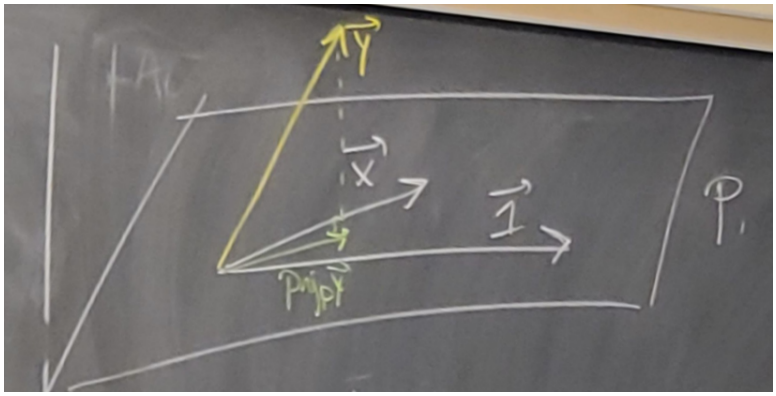
This is n equations with 2 variables, but we can rewrite as vector equations.

$$\vec{y} = m\vec{x} + b\vec{1} \quad \text{where } \vec{y} = y_1, \ldots, y_n, \vec{x} = x_1, \ldots, x_n, \text{ and } \vec{1} = \underbrace{(1, 1, \ldots, 1)}_{n \text{ components}}.$$

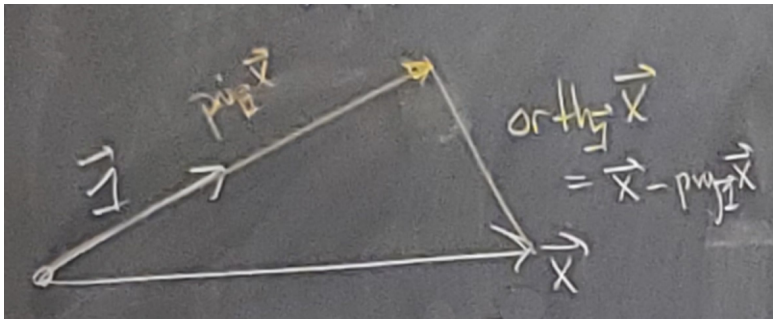
Trying to write $\vec{y}$ as a linear combination of $\vec{x}$ and $\vec{1}$. Consider the plane $P = \text{span}\{\vec{x}, \vec{1}\}$. Inconsistent implies that $\vec{y} \notin P$.

But the vector in P that is closest to $\vec{y}$ is $\text{proj}_P\, \vec{y}$. (Closest under Euclidean distance), recall $\langle u, v \rangle = u \cdot v$ in $\mathbb{R}^n$.

$$\text{dist}^2 = \sum (y_i - v_i)^2, \quad \vec{v} \in P = \underbrace{\vec{y} \cdot \vec{v} = \langle y, v \rangle}_{\text{"least squares"}}$$



Easiest way to project onto P is to have an orthogonal basis for P. We need an orthogonal basis for P:



$P = \operatorname{span}\{1, \operatorname{orth}_1 x\} = \operatorname{span}\{1, v\}$ where

$$\begin{aligned} v &= x - \operatorname{proj}_1 x \\ &= x - \frac{\langle x, 1\rangle}{\langle 1, 1\rangle} 1 \\ &= x - \frac{\langle x, 1\rangle}{n} 1 \end{aligned}$$

Then

$$\begin{aligned} \operatorname{proj}_P y &= \frac{\langle y, 1\rangle}{\langle 1, 1\rangle} 1 + \frac{\langle y, v\rangle}{\langle v, v\rangle} v \\ &= \frac{\langle y, 1\rangle}{n} 1 + \frac{\langle y, v\rangle}{\langle v, v\rangle} x - \frac{\langle y, v\rangle}{\langle v, v\rangle} \cdot \frac{\langle x, 1\rangle}{n} 1 \\ &= \underbrace{\frac{\langle y, v\rangle}{\langle v, v\rangle}}_{\substack{m \\ \text{(slope)}}} x + \underbrace{\left(\frac{\langle y, 1\rangle}{n} - \frac{\langle y, v\rangle}{\langle v, v\rangle} \cdot \frac{\langle x, 1\rangle}{n} \right)}_{\substack{b \\ (y\text{-intercept})}} 1 \end{aligned}$$

$$\begin{aligned}
\langle v, v\rangle &= \left\langle x - \frac{\langle x,1\rangle}{n}1,\ x - \frac{\langle x,1\rangle}{n}1 \right\rangle \\
&= \langle x,x\rangle - 2\left\langle x,\ \frac{\langle x,1\rangle}{n}1 \right\rangle + \left\langle \frac{\langle x,1\rangle}{n}1,\ \frac{\langle x,1\rangle}{n}1 \right\rangle \\
&= \langle x,x\rangle - 2\frac{\langle x,1\rangle}{n}\langle x,1\rangle + \frac{\langle x,1\rangle^2}{n^2}\langle 1,1\rangle \\
&= \langle x,x\rangle - 2\frac{\langle x,1\rangle^2}{n} + \frac{\langle x,1\rangle^2}{n} \\
&= \frac{n\langle x,x\rangle - \langle x,1\rangle^2}{n}
\end{aligned}$$

$$\begin{aligned}
\langle v, y\rangle &= \langle x,y\rangle - \frac{\langle x,1\rangle\langle y,1\rangle}{n} \\
&= n\langle x,y\rangle - \langle x,1\rangle\langle y,1\rangle
\end{aligned}$$

Then

$$\begin{aligned}
m &= \frac{\langle v,y\rangle}{\langle v,v\rangle} \\
&= \frac{n\langle x,y\rangle - \langle x,1\rangle\langle y,1\rangle}{n\langle x,x\rangle - \langle x,1\rangle^2}
\end{aligned}$$

Then

$$\begin{aligned}
b &= \frac{\langle y,1\rangle}{n} - m\frac{\langle x,1\rangle}{n} \\
&= \bar{y} - m\bar{x}
\end{aligned}$$

Which is what we got form the probabalistic approach.

Claim: The two-sets of formulas are identical

$$\begin{aligned}
m &= \frac{\mathrm{Cov}(X,Y)}{S_x^2} = \rho\frac{S_y}{S_x} \\
&= \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}
\end{aligned}$$

Useful formula:

$$\begin{aligned}
\sum(x_i-\bar{x})(y_i-\bar{y}) &= \sum(x_iy_i - x_i\bar{y} - y_i\bar{x} + \bar{x}\bar{y}) \\
&= \sum x_iy_i - \bar{y}\sum x_i - \bar{x}y_i + \bar{x}\bar{y}\sum 1 \\
&= \sum x_iy_i - \bar{y}n\cdot\frac{1}{n}\sum x_i - \bar{x}n\cdot\frac{1}{n}\sum y_i + n\bar{x}\bar{y} \\
&= \sum x_iy_i - n\bar{x}\bar{y} - n\bar{x}\bar{y} + n\bar{x}\bar{y} \\
&= \sum x_iy_i - n\bar{x}\bar{y}
\end{aligned}$$

$$
\begin{aligned}
m &= \frac{\operatorname{Cov}(X,Y)}{S_x^2} = \rho\frac{S_y}{S_x} \\
&= \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2} \\
&= \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} \\
&= \frac{\langle x,y\rangle - n\left(\frac{\langle x,1\rangle}{n}\right)\left(\frac{\langle y,1\rangle}{n}\right)}{\langle x,x\rangle - n\left(\frac{\langle x,1\rangle}{n}\right)^2} \\
&= \frac{n\langle x,y\rangle - \langle x,1\rangle\langle y,1\rangle}{n\langle x,x\rangle - \langle x,1\rangle^2} \\
&= \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{n\sum x_i^2 - (\sum x_i)^2} \\
&= m
\end{aligned}
$$

11.4 Properties of the Least-Squares Estimators: Simple Linear Regression

The Coefficents of the Best-Fit Line are Estimators. We have shown the best-fit line to be $\widehat{y} = \widehat{\beta_0} + \widehat{\beta_1}x$.

Where $\widehat{\beta_1} = \dfrac{S_{xy}}{S_{xx}}$ with $S_{xy} = \sum(x_i-\bar{x})(y_i-\bar{y})$ and $S_{xx} = \sum(x_i-\bar{x})^2$. (i.e. $\operatorname{Cov}(X,Y)$ and $\operatorname{Var}[X]$ via use of $\bar{x}$, $\bar{y}$ for μ_x, μ_y).

$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}x$

We recognize $\widehat{\beta_0}$ and $\widehat{\beta_1}$ as stats dependent upon $\bar{x}, S_x^2$, and $\bar{y}$.

But what exactly are they estimating? In theory, $X \times Y$ distributed via pdf $f(x,y)$ and there is some "best" linear relationship over the same probabilty space.

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

In other words, when we are "at" x_i,

$$\mathrm{E}[Y] = \beta_0 + \beta_1 x_i.$$

A modern approach to the distribution of Y at x is to add an error parameter ε. That is,

$$y = \underbrace{\beta_0 + \beta_1 x}_{\substack{\text{deterministic}\\ \text{component}\\ \text{of } Y}} + \underbrace{\varepsilon}_{\substack{\text{"random"}\\ \text{component}}}$$

Still want $\mathrm{E}[Y] = \beta_0 + \beta_1 x$. By linearity,

$$\begin{aligned}
\mathrm{E}[Y] &= \mathrm{E}[\beta_0 + \beta_1 x + \varepsilon] \\
&= \mathrm{E}[\beta_0 + \beta_1 x] + \mathrm{E}[\varepsilon] \\
&\implies \mathrm{E}[\varepsilon] = 0 \quad \text{error averages to zero}
\end{aligned}$$

We make the additional assumption that the variance of ε is independent of x. That is $\mathrm{Var}[Y] = \mathrm{Var}[\varepsilon] = \sigma^2$. The "value" of y *depends* on x, but the spread of the y-values does not.

Proposition: $\widehat{\beta_0}, \widehat{\beta_1}$ are unbiased estimators of β_0, β_1 where $Y = \beta_0 + \beta_1 x + \varepsilon$.

Reason:

$$\begin{aligned}
\mathrm{E}(\widehat{\beta_1}) &= \mathrm{E}\left[\frac{S_{xy}}{S_{xx}}\right] \\
&= \mathrm{E}\left[\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{S_{xx}}\right] \\
&= \mathrm{E}\left[\frac{\sum(x_i - \bar{x})y_i - \sum(x_i - \bar{x})\bar{y}}{S_{xx}}\right] \\
&= \mathrm{E}\left[\frac{\sum(x_i - \bar{x})y_i - \bar{y}\overbrace{\sum(x_i - \bar{x})}^{0 \text{ by def'n of } \bar{x}}}{S_{xx}}\right] \\
&= \mathrm{E}\left[\frac{\sum(x_i - \bar{x})y_i}{S_{xx}}\right] \\
&= \frac{\sum(x_i - \bar{x})\,\mathrm{E}(y_i)}{S_{xx}} \\
&= \frac{\sum(x_i - \bar{x})(\beta_0 + \beta_1 x_i)}{S_{xx}} \\
&= \frac{\overbrace{\beta_0 \sum(x_i - \bar{x})}^{0} + \beta_1 \sum(x_i - \bar{x})x_i}{S_{xx}} \\
&= \frac{\beta_1 \sum(x_i^2 - x_i\bar{x})}{S_{xx}} \\
&= \frac{\beta_1 (\sum x_i^2 - \bar{x}\sum x_i)}{S_{xx}} \\
&= \frac{\beta_1 (\sum x_i^2 - n\bar{x}^2)}{S_{xx}} \\
&= \beta_1 \frac{S_{xx}}{S_{xx}} \\
&= \beta_1
\end{aligned}$$

Therefore our estimator is unbiased.

The point of all that: $\mathrm{E}(\widehat{\beta_1}) = \widehat{\beta_1}$.

For $\mathrm{E}(\widehat{\beta_0}) = \mathrm{E}\left[\bar{y} - \widehat{\beta_1}\bar{x}\right] = \mathrm{E}\left[\bar{y}\right] - \bar{x}\,\mathrm{E}(\widehat{\beta_1})$.

But $\bar{y} = \dfrac{1}{n}\sum y_i = \dfrac{1}{n}\sum(\beta_0 + \beta_1 x + \varepsilon) = \beta_0 + \beta_1\bar{x} + \varepsilon$.

and $\mathrm{E}\left[\bar{y}\right] = \beta_0 + \beta_1\bar{x}$. and $\mathrm{E}\left[\widehat{\beta_0}\right] = \beta_0 + \beta_1\bar{x} - \beta_1\bar{x} = \beta_0$.

Corollary: $\mathrm{Var}(\widehat{\beta_1}) = \dfrac{\sigma^2}{S_{XX}}, \mathrm{Var}(\widehat{\beta_0}) = \dfrac{\sigma^2\sum x_i^2}{nS_{XX}}$.

Reason

$$\begin{aligned}
\mathrm{Var}(\widehat{\beta_1}) &= \mathrm{Var}\left(\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}\right) \\
&= \mathrm{Var}\left(\frac{\sum(x_i - \bar{x})^2 y_i}{S_{XX}}\right) \\
&= \frac{1}{{S_{XX}}^2}\,\mathrm{Var}\left(\sum(x_i - \bar{x})y_i\right) \\
&= \frac{\sum(x_i - \bar{x})^2\,\mathrm{Var}\left(Y_i\right)}{{S_{XX}}^2}
\end{aligned}$$

assumption: $\mathrm{Var}\left(Y\right) = \mathrm{Var}\left(\varepsilon\right) = \sigma^2$

$$\begin{aligned}
&= \frac{\sum(x_i - \bar{x})^2\sigma^2}{{S_{XX}}^2} \\
&= \frac{\sigma^2\sum(x_i - \bar{x})^2}{{S_{XX}}^2} \\
&= \frac{\sigma^2 S_{XX}}{{S_{XX}}^2} \\
&= \frac{\sigma^2}{S_{XX}}.
\end{aligned}$$

$$\mathrm{Var}(\widehat{\beta_0}) = \mathrm{Var}\left(\bar{y} - \widehat{\beta_1}\bar{x}\right)$$

Note in our ε set up, $\bar{y}$ and $\widehat{\beta_1}$ both functions of Y_i's... may be independent.

$$\begin{aligned}
&= \mathrm{Var}\left(\bar{y}\right) - \mathrm{Var}(\widehat{\beta_1}\bar{x}) - 2\,\mathrm{Cov}(\bar{y}, \widehat{\beta_1}\bar{x}) \\
&= \mathrm{Var}\left(\bar{y}\right) + \bar{x}^2\,\mathrm{Var}(\widehat{\beta_1}) - 2\bar{x}^2\,\mathrm{Cov}(\bar{y}, \widehat{\beta_1})
\end{aligned}$$

And i.) $\operatorname{Var}(\bar{y}) = \operatorname{Var}(\bar{\varepsilon}) = \frac{1}{n}\operatorname{Var}(\varepsilon) = \frac{1}{x}\sigma^2$. And ii.)

$$\begin{aligned}
\operatorname{Cov}(\bar{y}, \widehat{\beta_1}) &= \operatorname{Cov}\left(\frac{1}{n}\sum y_i,\ \sum \frac{(x_i - \bar{x})y_i}{S_{XX}}\right) \\
&= \sum \frac{(x_i - \bar{x})\operatorname{Var}(Y_i)}{nS_{xx}} + \sum\sum_{i\neq j} \frac{(x_j 0\bar{x})}{nS_{XX}} \underbrace{\operatorname{Cov}(Y_i, Y_j)}_{\text{independent } =0} \\
&= \frac{\sigma^2}{nS_{XX}} \sum (x_i - \bar{x}) \\
&= 0.
\end{aligned}$$

Then

$$\begin{aligned}
\operatorname{Var}(\widehat{\beta_0}) &= \operatorname{Var}(\bar{y}) + \bar{x}^2 \operatorname{Var}(\widehat{\beta_1}) - 2\bar{x}^2 \operatorname{Cov}(\bar{y}, \widehat{\beta_1}) \\
&= \frac{\sigma^2}{n} + \bar{x}^2 \cdot \frac{\sigma^2}{S_{XX}} \\
&= \sigma^2 \left(\frac{S_{XX} + n\bar{x}^2}{nS_{XX}}\right)
\end{aligned}$$

$$\begin{aligned}
S_{XX} &= \sum (x_i - \bar{x})^2 \\
&= \sum (\bar{x}^2 - 2\bar{x}) + \sum {x_i}^2 \\
&= n\bar{x}^2 - 2\bar{x}\sum x_i \\
&= n\bar{x}^2 - 2\bar{x}n\left(\frac{\sum x_i}{n}\right) + \sum {x_i}^2 \\
&= \sum {x_i}^2 - n\bar{x}^2 \qquad \text{i.e. } n\bar{x}^2 = \sum {x_i}^2 - S_{XX}
\end{aligned}$$

$$\begin{aligned}
\operatorname{Var}(\widehat{\beta_0}) &= \sigma^2 \left(\frac{S_{XX} + \sum {x_i}^2 - S_{XX}}{nS_{XX}}\right) \\
&= \frac{\sigma^2 \sum {x_i}^2}{nS_{XX}}
\end{aligned}$$

Corollary: $\operatorname{Cov}(\widehat{\beta_0}, \widehat{\beta_1}) = -\frac{\bar{x}\sigma^2}{S_{XX}}$. Proof in textbook. Note: $\widehat{\beta_0}, \widehat{\beta_1}$ guaranteed to be dependent when $\bar{x} = 0$.

Topic: estimating σ^2.

We have working with the assumption that $\operatorname{Var}(y) = \operatorname{Var}(\varepsilon) = \sigma^2$, but this is usually unknown.

In the past, we used $\operatorname{Var}(Y) = \frac{1}{n-1}\sum (y_i - \bar{y})^2$.

In our new setup, our point estimator for y_i is no longer $\bar{y}$. Our estimator of Y_i is the best-fit line: $\operatorname{E}(Y_i) = \widehat{\beta_0} + \widehat{\beta_1} x$.

We define S^2 via sum square error

$$\text{SSE} = \sum_{i=1}^{n} (y_i, \widehat{y_i})^2.$$

The term inside the summation is the residual; it is the error between the prediction and observed.

We need to define an unbiased estimator for σ^2. Let

$$\begin{aligned} \widehat{\theta} &:= K \cdot \text{SSE} \\ &= K \sum (y_i - (\widehat{\beta_0} + \widehat{\beta_1} x_i))^2 \end{aligned}$$

By computing $\text{E}(\widehat{\theta})$, we choose K so that $\text{E}(\widehat{\theta}) = \sigma^2$.

As we did in chapter 9, we can show that $\text{E}(\text{SSE}) = (n-2)\sigma^2$. So,

$$\boxed{S^2 = \frac{1}{n-2} \text{SSE} = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \widehat{y_i})^2}$$

Proposition: $\text{E}(S^2) = \sigma^2$, (unbiased estimator).
(aside: It would be nice to have a computation corollary like we did, i.e. $\text{Var}\,(x) = \text{E}\,(x^2) - \text{E}\,(x)^2$)

Corollary: The computation corollary

$$\text{SSE} = S_{YY} - \widehat{\beta_1} S_{XY}$$

$$\text{where} \quad S_{YY} = \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

Example: (11.16)
Potnecy of antibiotic (y): $(38, 43, 29), (32, 26, 33), (19, 27, 27), (14, 19, 21)$
After storage at x°F: $30, 50, 70, 90$

$$\bar{y} = 27, \qquad \bar{x} = 60$$

$$S_{xy} = \sum (x_i - \bar{x})(y_i i \bar{y}) = -1900$$

$$S_{xx} = \sum (x_i - \bar{x})^2 = 6000$$

$$S_{yy} = \sum (y_i - \bar{y})^2 = 792$$

Then $\widehat{\beta_1} = \dfrac{S_{xy}}{S_{xx}} = -\dfrac{19}{60}$ and

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1} \bar{X} x = 27 - \left(\frac{-19}{60} \right) 60 = 46.$$

So the best fit line is $\widehat{y} = 46 - \frac{19}{60}x$. For

$$\begin{aligned} S^2 &= \frac{1}{n-2}\,\mathrm{SSE} \\ &= \frac{1}{n-2}\left(S_{yy} - \widehat{\beta}_1 S_{xy}\right) \\ &= \frac{1}{12-2}\left(792 - \left(\frac{-19}{60}\right)(-1900)\right) = \frac{571}{30} \\ &\approx 19.0\overline{3} \end{aligned}$$

Discussion: Do we know how S^2 is distributed? This depends on our assumption on how ε is distributed.

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

We have $\mathrm{E}(\varepsilon) = 0$ and $\mathrm{Var}(\varepsilon) = \sigma^2$, and thus are independent of X.

Note that the distribution of the point estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are σ^2 dependent (which we will estimate with S^2).

However, the common assumption is that "everything" is normal! We assume $\varepsilon \sim \mathrm{N}(0, \sigma^2)$.

We can now prove a Fisher's Theorem like result for the new S^2.

Namely... Theorem: $\dfrac{(n-2)S^2}{\sigma^2} = \dfrac{\mathrm{SSE}}{\sigma^2} \sim \chi^2(n-2)$, moreover S^2 is independent of $\widehat{\beta}_0$ and $\widehat{\beta}_1$. Proof: Omitted for mental health reasons.

11.5 Inferences concerning the point estimators

Under the assumption $\varepsilon \sim \mathrm{N}(0, \sigma^2)$, we have that $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are normal $\mathrm{N}(\beta_i, \mathrm{Var}(\beta_i))$. Thus we can do confidence intervals for the true $\widehat{\beta}_0$ and $\widehat{\beta}_1$.

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

Use the same $\mathcal{Z}/\mathcal{T}$ rules as before, ($n \leq 30$ use $\mathcal{T}$ and $n > 30$ use $\mathcal{Z}$).

As always, the standardization is

$$\frac{\widehat{\theta} - \theta}{\sqrt{\sigma_{\widehat{\theta}}}} = \frac{\widehat{\theta} - \theta}{\sqrt{\mathrm{Var}(\widehat{\theta})}}$$

Example: Given $Z_{\alpha/2}$ ($n > 30$), $\widehat{\beta}_1 \pm Z_{\alpha/2}\sqrt{\mathrm{Var}(\widehat{\beta}_1)}$ is an α-level 2-sided C.I. for the true β_i.

Example: Find a 90% C.I. for the slope $\widehat{\beta}_1$ of the potency example. We have $\widehat{\beta}_1 = \dfrac{-19}{60} \approx$

$-0.3166\overline{7}$.

$$\begin{aligned}\text{Var}(\widehat{\beta_1}) &= \frac{\sigma^2}{S_{xx}} \\ &\approx \frac{S^2}{S_{xx}} \\ &= \frac{19.03}{6000} \\ &= 0.00317 \\ \implies S_{\widehat{\beta_1}} &= \sqrt{\text{Var}(\widehat{\beta_1})} = 0.0563\end{aligned}$$

Using S^2 implies we use S^2 degrees of freedom, $n = 12$ implies 10 degrees of freedom. Thus $S^2 \sim \chi^2(10)$. We need $t_{0.05}(10) = 1.812$ by table and

$$\begin{aligned}&\widehat{\beta_1} \pm t_{0.05}(10) S_{\widehat{\beta_1}} \\ &-0.31667 \pm (1.812)(0.0563) \\ &-0.31667 \pm 0.1020 \\ &(-0.4187,\ -0.214)\end{aligned}$$

Example: (Potency again)
We will assume that it is commonly know that penicillin derivatives are considered "effective" if when stored in a deep freezer (0°F) the potency is 50 or better. We wish to test if our new drug is "effective".

We have $\widehat{y} = \widehat{\beta_0} + \widehat{\beta_1} x = 46 - \frac{19}{60}x$. When $\widehat{y}(0) = \widehat{\beta_0} = 46$. We construct the test: If the true $\widehat{\beta_0} = 50$, given our sample observation $\widehat{\beta_0} = 46\ldots$ is this rare or not.

$$H_0 : \widehat{\beta_0} = 50 \qquad H_a : \widehat{\beta_0} < 50$$

Again, $S^2 \sim \chi^2(10)$,... we need to use t-test. We are assuming $\widehat{\beta_0} \sim \text{N}(\beta_0, \text{Var}(\beta_0)) \approx \text{N}(\beta_0, \text{Var}(\widehat{\beta_0}))$. p-value:

$$\begin{aligned}p &= \Pr\left(\widehat{\beta_0} \le 46 \mid \widehat{\beta_0} = 50\right) \\ &= \Pr\left(\frac{\widehat{\beta_0} - \beta_0}{\sqrt{\text{Var}(\widehat{\beta_0})}} \le \frac{46 - 50}{\sqrt{\text{Var}(\widehat{\beta_0})}}\right)\end{aligned}$$

Pausing to compute some of these

$$\text{Var}(\widehat{\beta_0}) = \frac{\sigma^2 \sum {x_i}^2}{n S_{xx}} \sim \frac{S^2 \sum {x_i}^2}{n S_{xx}}$$

Here $S^2 = 19.0\overline{3}$, $n = 12$, and $S_{xx} = 6000$.

$\sum {x_i}^2 = 3 \cdot (30^2 + 50^2 + 70^2 + 90^2) = 49200$ and $\text{Var}(\widehat{\beta_0}) = \dfrac{19.03 \cdot 49200}{12 \cdot 6000} = 13.004$ and $\sqrt{\text{Var}(\widehat{\beta_0})} = 3.606$. Then,

$$p = \Pr\left(T \le \frac{46 - 50}{3.606}\right) = P(T \le -1.09) = 0.150641$$

Cannot reject the null hypothesis! So maybe it *is* actually 50. Let's sell it!

11.6 Predictions via least squares regression line

Using $\widehat{y} = \widehat{\beta_0} + \widehat{\beta_1} x$. Our emphasis so far has been on the coefficients $\widehat{\beta_0}$ and $\widehat{\beta_1}$ and their distributions. The goal is to understand y as a function of x. y is the exact, deterministic expectation whereas $\widehat{y}$ is our best attempt to approximate something that isn't exact.

Example: (Potency Example Again)

$$\widehat{y} = \widehat{\beta_0} + \widehat{\beta_1} x = 46 - \frac{19}{60} x$$

If we store the antibiotic at temperature $x = 20°\text{F}$, what would we expect the potency to be?

$$\widehat{y} = 46 - \frac{19}{60}(20) = 39.\overline{6}$$

But, what does this really mean? This is the expected value of y when $x = 20$. If we store a bunch of samples at 20°F, we would expect the sample mean to be approximately $39.\overline{6}$. In other words, $\widehat{y} = \widehat{\beta_0} + \widehat{\beta_1} x$ is a point estimator of $\mathrm{E}(Y)$.

$\widehat{y}$ is a statistic of its own. And, where there exists point estimators, there exists interval estimators (confidence intervals). We have that

$$\widehat{\beta_i} \sim \mathrm{N}(\beta_i, \mathrm{Var}[\beta_i]).$$

So, $\widehat{y}$ is also normal for every fixed value of x.

To avoid formula confusion, we use x^* for the fixed x. Note:

$$\mathrm{E}(\widehat{y}(x^*)) = \mathrm{E}\left(\widehat{\beta_0} + \widehat{\beta_1} x^*\right) = \mathrm{E}\left(\widehat{\beta_0}\right) + x^* \mathrm{E}\left(\widehat{\beta_1}\right) = \beta_0 = \beta_1 x^*.$$

We now need the variance, $\mathrm{Var}[\widehat{y}]$.

$$\begin{aligned}
\mathrm{Var}[\widehat{y}] &= \mathrm{Var}\left[\widehat{\beta_0} + \widehat{\beta_1} x^*\right] \\
&= \mathrm{Var}\left[\widehat{\beta_0}\right] + (x^*)^2 \mathrm{Var}\left[\widehat{\beta_1}\right] + 2x^* \mathrm{Cov}(\widehat{\beta_0}, \widehat{\beta_1}) \\
&= \frac{\sigma^2 \sum {x_i}^2}{n S_{xx}} + (x^*)^2 \frac{\sigma^2}{S_{xx}} + 2x^* \left(\frac{-\bar{x}\sigma^2}{S_{xx}}\right) \\
&= \frac{\sigma^2}{S_{xx}} \left(\frac{\sum {x_i}^2 + n(x^*)^2 - 2x^* \bar{x} n}{n}\right) \\
\text{Note: } S_{xx} &= \sum {x_i}^2 - n\bar{x}^2 \\
&= \frac{\sigma^2}{S_{xx}} \left(\frac{S_{xx} + n\bar{x}^2 + n(x^*)^2 - 2x^* \bar{x} n}{n}\right) \\
&= \frac{\sigma^2}{S_{xx}} \left(\frac{S_{xx} + n\left[\bar{x}^2 + (x^*)^2 - 2x^* \bar{x}\right]}{n}\right) \\
&= \sigma^2 \left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right)
\end{aligned}$$

When estimating σ^2, make sure to use the new one:

$$S^2 = \frac{1}{n-2} \mathrm{SSE}.$$

Recap: $\widehat{y}$ estimator:

$$\mathrm{E}\left(\widehat{y}\right) = \beta_0 + \beta_1 x^*$$

and

$$\mathrm{Var}\left[\,\widehat{y}\,\right] = \sigma^2\left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right).$$

Standardize (as always):

$$\frac{\widehat{\theta} - \theta}{\sigma_{\widehat{\theta}}} = \frac{\widehat{y}(x^*) - (\beta_0 + \beta_1 x^*)}{S\sqrt{\dfrac{1}{n} + \dfrac{(x^* - x)^2}{S_{xx}}}}.$$

This is a test statistic with $(n-2)$ degrees of freedom (df).

It follows that $\alpha(1-\alpha)$ level confidence interval is given by:

$$\left(\widehat{\beta}_0 + \widehat{\beta}_1 x^*\right) \pm t_{\alpha/2}(n-2)S\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}.$$

Example: Find a 90% confidence interval for the average potency when stored at 20°F.

$x^* = 20$

$\widehat{y}(20) = 39.\overline{6}$ $\qquad\qquad$ $n = 12 \implies so\,\mathrm{d.f.} = 10$

$\bar{x} = 60$

$S^2 = \dfrac{571}{30}$ $\qquad\qquad$ $S_{xx} = 6000$

Answer:

$$39.\overline{6} \pm \underbrace{(1.812)}_{\text{table}}\sqrt{\underbrace{\frac{571}{30}}_{S^2}\left(\frac{1}{12} + \frac{(20-60)^2}{6000}\right)}$$

$$39.\overline{6} \pm 4.677$$

Remark: We can now talk about hypothesis testing.

For example,

$$H_0 : y(x^*) = y_0 \qquad \text{or} \qquad H_a : y(x^*) \neq y_0$$

Our t-stat

$$\mathcal{T} = \frac{y - 0 - (\widehat{\beta}_0 + \widehat{\beta}_1 x^*)}{S\sqrt{\dfrac{1}{n} + \dfrac{(x^* - \bar{x})^2}{S_{xx}}}}$$

11.7 Predictions on y

The difference between this section and the previous is perspective. In §11.6 we focused on the spread of the average of y. In other words, $\mathrm{E}(\widehat{y})$.

Wha tif instead, we wanted to focus on the distribution of the values that y can take when $x = x^*$. Recall:

$$y = \underbrace{\beta_0 + \beta_1 x}_{\text{deterministic}} + \underbrace{\varepsilon}_{\text{“spread” of } y}$$

and $\mathrm{Var}(y) = \mathrm{Var}(\varepsilon) = \sigma^2$. Our estimator for y at x^* is defined

$$\widehat{y}^* = \widehat{\beta}_0 + \widehat{\beta}_1 x^* + \varepsilon.$$

Since $\varepsilon \sim \mathrm{N}(0, \sigma^2)$.

$$\begin{aligned}
\mathrm{E}(\widehat{y}^*) &= \mathrm{E}\left(\widehat{\beta}_0 + \widehat{\beta}_1 x^*\right) + \mathrm{E}(\varepsilon) \\
&= \mathrm{E}(\widehat{y}(x^*)) \\
&= \beta_0 + \beta_1 x^*
\end{aligned}$$

Computing variance is surprisingly easy. As before, $\widehat{y}^*$ is a sum of Normal random variables. Hence, $\widehat{y}^*$ is Normally distributed.

Moreover, our assumption on ε is that it is independent of x.

We have that S^2 is independent of $\widehat{\beta}_0$ and $\widehat{\beta}_1$. So

$$\begin{aligned}
\mathrm{Var}(\widehat{y}^*) &= \mathrm{Var}\left(\widehat{\beta}_0 + \widehat{\beta}_1 x^*\right) + \mathrm{Var}(\varepsilon) && \text{(by independence)} \\
&= \sigma^2 \left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_x x}\right) + \sigma^2 \\
&= \sigma^2 \left(1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_x x}\right)
\end{aligned}$$

This is the only difference from last sections $\mathrm{Var}(\widehat{y})$.

When σ^2 unknown.

$$\mathrm{Var}(\widehat{y}^*) = S^2 \left(1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_x x}\right)$$

For the $1 - \alpha$ level confidence interval (two sided)

$$\Pr\left(|T| > t_{\alpha/2}(\mathrm{df})\right) = 1 - \alpha$$

$$\mathcal{Z} = \frac{y^* - \mathrm{E}(\widehat{y}^*)}{\sigma_{\widehat{y}^*}}$$

or

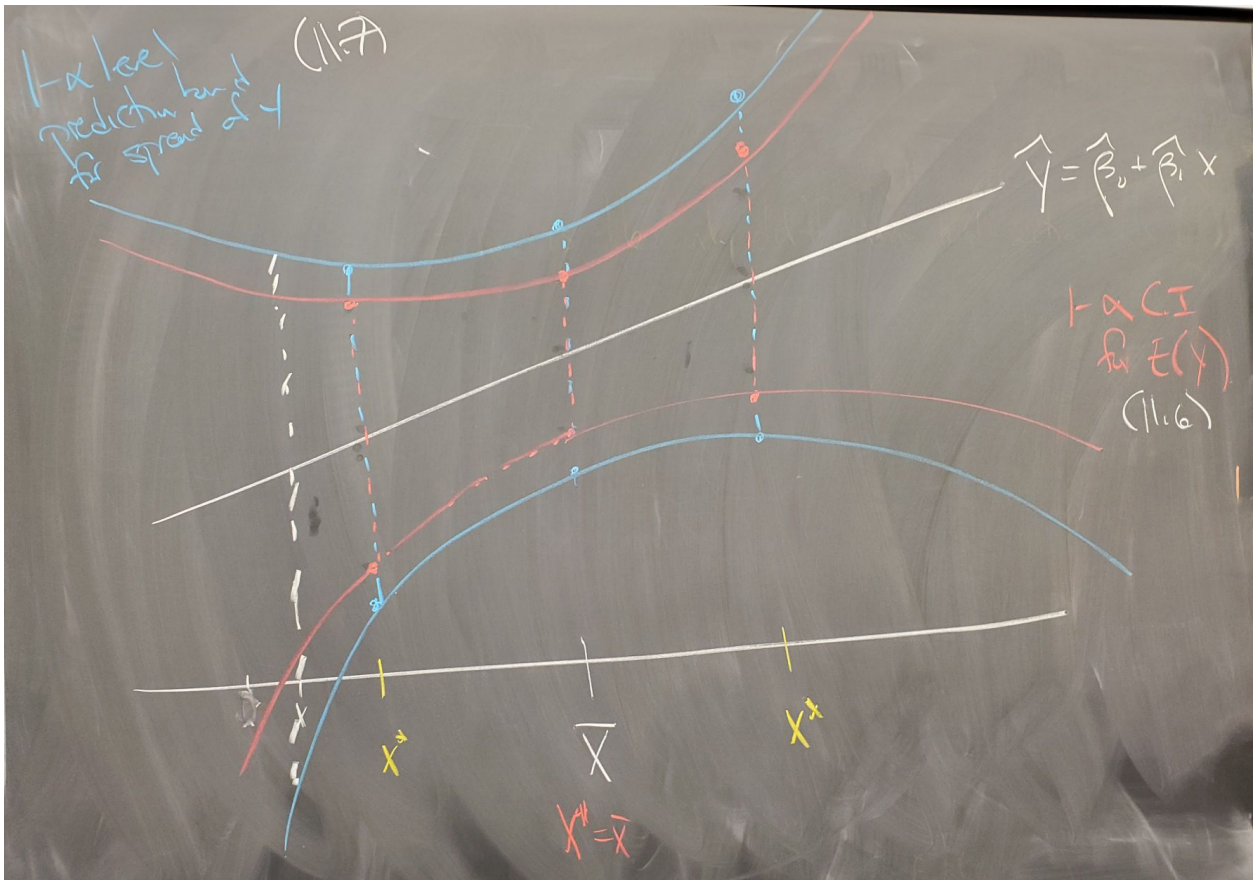
$$\mathcal{T} = \frac{y^* - \left(\widehat{\beta}_0 + \widehat{\beta}_1 x^*\right)}{S\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_x x}}}$$

t-distributed with $n-2$ degrees of freedom and $1-\alpha$ level confidence interval is

$$\left(\widehat{\beta}_0 + \widehat{\beta}_1 x^*\right) \pm t_{\alpha/2}(\text{df}) S \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_x x}}$$

Discussion: Confidence and prediction bands.
In §11.6, we had C.I. for the mean of $\widehat{y}(x^*)$.
In §11.7, we have C.I. for the spread of y at x^*,

In either case, as x^* moves away from $\bar{x}$, the $\dfrac{(x^* - \bar{x})^2}{S_{xx}}$ increases as does the standard error. i.e. the intervals get longer.



Example: Potency of antibiotic when stored at 20°F. Last day we showed that $\mathrm{E}\,(\widehat{y}(20))$ had 90% confidence interval

$$\underbrace{39.6 \pm 4.677}_{\text{red cross-section}}$$

What spread of potency would we expect to see with 90% confidence if stored at 20°F. The same calculation as the last but with new standard error

$$39.\overline{6} \pm 1.812 \sqrt{1 + \frac{1}{12} + \frac{1600}{6000}}$$

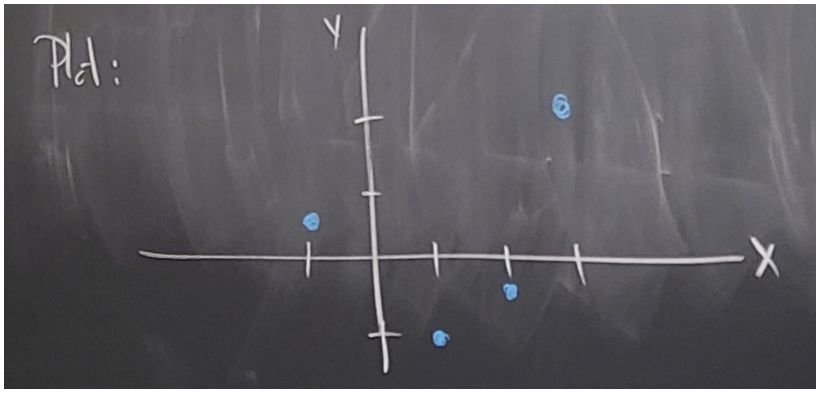
$$\underbrace{39.6 \pm 5.069}_{\text{blue cross-section}}$$

11.10 Multiple Linear Regression

Example: Consider the data:

x	-1	1	2	3
y	0.5	-1	-0.5	2

The the plot does not look linear:



Seems unlikely that a line will fit well. What about a parabola? Want $y = \beta_0 + \beta_1 x + \beta_2 x^2$. Of course, in the probabilistic setup:

$$y = \underbrace{\beta_0 + \beta_1 x + \beta_2 x^2}_{\text{deterministic}} + \underbrace{\varepsilon}_{\text{error}}$$

Using the data:

$$\begin{aligned} \frac{1}{2} &= \beta_0 + \beta_1(-1) + \beta_2(-1)^2 \\ -1 &= \beta_0 + \beta_1(1) + \beta_2(1)^2 \\ -\frac{1}{2} &= \beta_0 + \beta_1(2) + \beta_2(2)^2 \\ 2 &= \beta_0 + \beta_1(3) + \beta_2(3)^2 \end{aligned}$$

But this is the same idea as before, just in higher dimensions (3 coefficents).

$$\vec{y} = \beta_0 \vec{1} + \beta_1 \vec{x} + \beta_2 \vec{x^2}$$

$$= \beta_0 \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + \beta_1 \begin{pmatrix} -1 \\ 1 \\ 2 \\ 3 \end{pmatrix} + \beta_2 \begin{pmatrix} 1 \\ 1 \\ 4 \\ 9 \end{pmatrix}$$

Or, as a matrix equation,

$$\begin{bmatrix} 1 & -1 & 1 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 1/2 \\ -1 \\ -1/2 \\ 2 \end{bmatrix}.$$

Remark: This is why this process is called **linear** regression.

The system for the coefficients will be a linear system. It has nothingto do with the underlying model... WHich here is quadratic $y = \beta_0 + \beta_1 x + \beta_2 x^2$.

Just as before, we have an overdetermined system, which implies there is no solution. Hence, the best we can do is find the projection.

We need a way to solve this projection problem in general.

Topic: The Normal Equations:
In complete generality,

$$y = \beta_0 + \beta_1 f_1(x) + \beta_2 f_2(x) + \cdots + \beta_n f_n(x) + \varepsilon.$$

Let there be m data points (x_i, y_i).

This yields a matrix equation:

$$\begin{bmatrix} 1 & f_1(x_1) & f_2(x_1) & \cdots & f_n(x_1) \\ 1 & f_1(x_2) & f_2(x_2) & \cdots & f_n(x_2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & f_1(x_n) & f_2(x_n) & \cdots & f_n(x_n) \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Write $\mathbf{X}\vec{\beta} = \vec{y}$. Then $\mathbf{X}$ is mxn.

Note, this setup really only makes sense when the system is overdetermined (more equations than unknowns). That is, more rows then columns in $\mathbf{X}$. i.e. $m > n$. We seek the coefficient vectors $\widehat{\beta} = [\widehat{\beta}_1, \ldots, \widehat{\beta}_n]$ such that $\left\| y - \mathbf{X}\widehat{\beta} \right\|$ is minimized (LSR).

In the gernal case, there are 2 new issues to content with:

① no reason the column vectors of $\mathbf{X}$ are actually a basis for $\operatorname{Col}\mathbf{X}(mightjustbeaspanningset)$.

② Certainly no reason columns are an orthogonal set.

The implication of ① is that a least squares solution $\widehat{\beta}$ need not be unique. As for ②, ends we don't need this. By definition $\vec{y} - \mathbf{X}\widehat{\beta}$ is orthogonal to $\operatorname{Col}\mathbf{X}$.

In other words, $\vec{y} - \mathbf{X}\widehat{\beta}$ must lie in the null space of $\mathbf{X}^T$. (Recall $(\text{Col})^\perp = \text{null}\, A^T$).

Thus, $\mathbf{X}^T(y - \mathbf{X}\widehat{\beta}) = \vec{0}$.

$$\implies \underbrace{\mathbf{X}^T y}_{m \text{ vector}} - \underbrace{\mathbf{X}^T\mathbf{X}\widehat{\beta}}_{m \text{ vector}} = \underbrace{0}_{m}$$

$$\implies \mathbf{X}^T\mathbf{X}\widehat{\beta} = \mathbf{X}^T y$$

Definition: The normal equations:

$$\mathbf{X}^T\mathbf{X}\widehat{\beta} = \mathbf{X}^T y$$

Facts:

① $\mathbf{X}^T\mathbf{X}$ is an nxn matrix.

② The normal equation is always a consistent system. That is, there is a solution to the least squares problem.

③ If the columns of $\mathbf{X}$ are linearly independent, then $\mathbf{X}^T\mathbf{X}$ is an invertible square matrix and the unique solution to the least squares problem is

$$\widehat{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T y$$

Example: Our best fit parabola.

$$\mathbf{X} = \begin{bmatrix} 1 & -1 & 1 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \end{bmatrix}, \qquad \mathbf{X}^T = \begin{bmatrix} 1 & 1 & 1 & 1 \\ -1 & 1 & 2 & 3 \\ 1 & 1 & 4 & 9 \end{bmatrix}$$

$$\widehat{\beta} = [\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2], \quad y = \left[\frac{1}{2}, -1, -\frac{1}{2}, 2\right]$$

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 4 & 5 & 15 \\ 5 & 15 & 35 \\ 15 & 35 & 99 \end{bmatrix}$$

Note that $\det(\mathbf{X}^T\mathbf{X}) = 440 \neq 0$ so it is invertible.

$$\begin{aligned} \widehat{\beta} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y \\ &= \begin{bmatrix} -41/44 \\ -379/440 \\ 53/88 \end{bmatrix} \end{aligned}$$

So, the best-fit parabola is

$$\begin{aligned} y &= -\frac{41}{44} - \frac{379}{440}x + \frac{53}{88}x^2 \\ &\approx -0.932 - 0.861x + 0.602x^2 \end{aligned}$$

Discussion: Revisit the best-fit line

Given n (x_i, y_i)'s, $y = \beta_0 + \beta_1 x + \varepsilon$.

$$\Longrightarrow \begin{bmatrix} \vec{1} & \bar{x} \end{bmatrix} \widehat{\beta} = \vec{y} \qquad \text{and} \qquad \widehat{\beta} = \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{bmatrix}$$

Here $\mathbf{X} = \begin{bmatrix} \vec{1} & \bar{x} \end{bmatrix}$ is a $n \times 2$ matrix and $\mathbf{X}^T = \begin{bmatrix} 1^T \\ x^T \end{bmatrix}$ is $2 \times n$.

Then

$$\begin{aligned} \mathbf{X}^T\mathbf{X} &= \begin{bmatrix} 1^T \\ x^T \end{bmatrix} \begin{bmatrix} \vec{1} & \bar{x} \end{bmatrix} \\ &= \begin{bmatrix} 1 \cdot 1 & 1 \cdot x \\ x \cdot 1 & x \cdot x \end{bmatrix} \\ &= \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum {x_i}^2 \end{bmatrix} \end{aligned}$$

Normal equation $X^T X\widehat{\beta} = X^T y$. Can we solve uniquely?

$$\begin{aligned}\det\big(X^T X\big) &= n\sum {x_i}^2 - \Big(\sum x_i\Big)^2 \\ &= n\sum {x_i}^2 - n^2\bar{x}^2 \\ &= n\Big(\sum {x_i}^2 - n\bar{x}\Big) \\ &= nS_{xx} \\ &> 0\end{aligned}$$

Therefore $X^T X$ is invertible! ($\det A \neq 0 \iff A^{-1}$ exists)

So $\widehat{\beta} = (X^T X)^{-1} X^T y$

Recall: $\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \dfrac{1}{\det A}\big[(\operatorname{tr} A)I - A\big] = \dfrac{1}{ad-bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$.

So

$$(X^T X)^{-1} = \frac{1}{nS_{xx}}\begin{bmatrix} \sum {x_i}^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix}$$

and

$$X^T y = \begin{bmatrix} 1^T \\ x^T \end{bmatrix} y = \begin{bmatrix} 1 \cdot y \\ x \cdot y \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

Then

$$\begin{aligned}\widehat{\beta} &= \frac{1}{nS_{xx}}\begin{bmatrix} \sum {x_i}^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix}\begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix} \\ &= \frac{1}{nS_{xx}}\begin{bmatrix} (\sum {x_i}^2)(\sum y_i) - (\sum x_i)(\sum x_i y_i) \\ -(\sum x_i)(\sum y_i) + n\sum x_i y_i \end{bmatrix}\end{aligned}$$

Same $\widehat{\beta}_0, \widehat{\beta}_1$ as earlier. But wait there's more..., Look at $(X^T X)^{-1}$:

$$\begin{aligned}(X^T X)^{-1} &= \frac{1}{nS_{xx}}\begin{bmatrix} \sum {x_i}^2 & -n\bar{x} \\ -n\bar{x} & n \end{bmatrix} \\ &= \begin{bmatrix} \dfrac{\sum {x_i}^2}{nS_{xx}} & -\dfrac{\bar{x}}{S_{xx}} \\ -\dfrac{\bar{x}}{S_{xx}} & \dfrac{1}{S_{xx}} \end{bmatrix}\end{aligned}$$

And

$$\sigma^2 (X^T X)^{-1} = \begin{bmatrix} \operatorname{Var}\Big(\widehat{\beta}_0\Big) & \operatorname{Cov}(\widehat{\beta}_0, \widehat{\beta}_1) \\ \operatorname{Cov}(\widehat{\beta}_0, \widehat{\beta}_1) & \operatorname{Var}\Big(\widehat{\beta}_1\Big) \end{bmatrix}$$

FACT: This always happens. $\sigma^2 (X^T X)^{-1}$ is the table of covariances:

$$\sigma^2 - a_{i\,j} = \operatorname{Cov}(\widehat{\beta}_i, \widehat{\beta}_j) \quad \text{where} \quad [a_{i\,j}] = (X^T X)^{-1}$$

Lastly,

$$S^2 = \frac{\text{SSE}}{n-2} = \frac{\sum (y_i - \widehat{y}_i)^2}{n-2}$$

And "by some matrix algebra" (Wackerly),

$$\text{SSE} = y^T y - \widehat{\beta}^T \mathbf{X}^T y.$$

Example: (Antibiotic again)

x	30	30	30	50	50	50	70	70	70	90	90	90
y	38	43	29	32	26	33	19	24	23	14	19	21

$n = 12$ and $X^T X = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum {x_i}^2 \end{bmatrix} = \begin{bmatrix} 12 & 720 \\ 720 & 49{,}200 \end{bmatrix}$. Then $\det(X^T X) = 72{,}000$. Then

$$(X^T X)^{-1} = \frac{1}{72{,}000} \begin{bmatrix} 49{,}200 & -700 \\ -700 & 12 \end{bmatrix}.$$

Note: $\text{Var}(\widehat{\beta}_0) = \frac{41}{60}\sigma^2$, and $\text{Var}(\widehat{\beta}_1) = \frac{\sigma^2}{6000}$.

Then $X^T y = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix} = \begin{bmatrix} 324 \\ 17{,}540 \end{bmatrix}$ and $\widehat{\beta} = (X^T X)^{-1} X^T y = \begin{bmatrix} 46 \\ -19/60 \end{bmatrix}$ (same as before). Hence

$$\widehat{y} = 46 - \frac{19}{60} x$$

$$\text{SSE} = y^T y - \widehat{\beta}^T X^T y = \frac{2961}{15}$$

$$\sigma^2 \approx \frac{\text{SSE}}{n-2} = \frac{2971/15}{10} \approx 19.8067.$$

11.11 A big ol theorem

Theorem: Let $Y_i = \beta_0 + \beta_1 f_1(x_i) + \cdots + \beta_k f_k(x_i) + \varepsilon_i$ where $\varepsilon_i \sim N(0, \sigma^2)$ (with $\text{E}[\varepsilon] = 0$). Then the least-squares estimates are given by $\widehat{\beta} = (X^T X)^{-1} X^T y$ provided $(X^T X)^{-1}$ exists. Then

① $\text{E}(\widehat{\beta}_i) = \widehat{\beta}_i$ (unbiased)

② $\text{Var}(\widehat{\beta}_1) = c_{ii}\sigma^2$ where $(X^T X)^{-1} = [c_{ij}]$

③ $\text{Cov}(\widehat{\beta}_i, \widehat{\beta}_j) = c_{ij}\sigma^2$

④ $S^2 = \frac{\text{SSE}}{n-(k+1)}$ with $k+1$ $\widehat{\beta}_i$'s and $\text{SSE} = y^T y - \widehat{\beta}^T X^T y$ and $\text{E}(S^2) = \sigma^2$ (unbiased)

⑤ $\widehat{\beta}_i$ is normally distributed.

⑥ $\frac{(n-(k+1))S^2}{\sigma^2} \sim \chi^2(n-(k+1))$

⑦ S^2 and $\widehat{\beta}_1$ are independent for all i.

11.12 Hypothesis Testing C.I.

Motivation: Testing a specific $\widehat{\beta}_i$.

We have $\widehat{\beta} = (X^T X)^{-1} T y = [\widehat{\beta}_0, \ldots, \widehat{\beta}_k]$. To pick off a specific $\widehat{\beta}_i$, we use the dot product.

Let e_i (standard basis vector), then $\widehat{\beta}_i = \underbrace{e_i \cdot \widehat{\beta}}_{\text{dot product}} = \underbrace{{e_i}^T \widehat{\beta}}_{\substack{\text{matrix} \\ \text{multiplication}}}$

Note: $\mathrm{E}(e_i \cdot \widehat{\beta}) = \mathrm{E}(1 \cdot \widehat{\beta}_i) = \mathrm{E}(\widehat{\beta}_i) = \beta_i$.
Same with $\mathrm{Var}(e_i \cdot \widehat{\beta}) = \mathrm{Var}(\widehat{\beta}_i)$.

So a test of the form

$$H_0 : \widehat{\beta}_i = (\beta_i)_0$$
$$H_a : \widehat{\beta}_i \neq (\beta_i)_0$$

yields same old statistics.

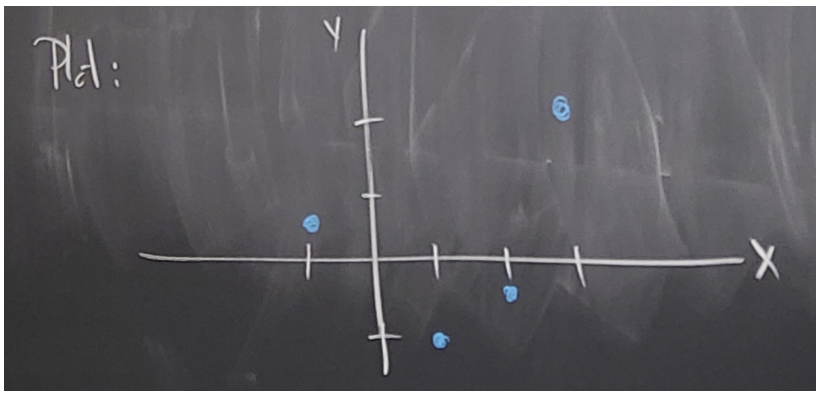
$$Z = \frac{\widehat{\beta}_i - (\beta_i)_0}{\sqrt{\mathrm{Var}(\beta_i)}} = \frac{\widehat{\beta}_i - (\beta_i)_0}{\sqrt{c_{i\,i}\sigma^2}} \qquad \text{with} \qquad [c_{i\,i}] = (X^T X)^{-1}$$

$$T = \frac{\widehat{\beta}_i - (\beta_i)_0}{S\sqrt{c_{i\,i}}} \qquad \text{with} \qquad n - (k+1) \text{ df.}$$

Example: We fit

x	-1	1	2	3
y	0.5	-1	-0.5	2

with $y = \beta_0 + \beta_1 x + \beta_2 x^2$ because it "looked" non-linear.



Is there evidence that the data is in fact non-linear? That is, are we reasonably certain $\beta_\alpha \neq 0$.

$$H_0 : \beta_\alpha = 0$$
$$H_a : \beta_\alpha \neq 0$$

We had

$$\begin{aligned}\widehat{\beta} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y \\ &= \begin{bmatrix} -41/44 \\ -379/440 \\ 53/88 \end{bmatrix} \\ &= \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{bmatrix}\end{aligned}$$

$e_3 = (0, 0, 1)$ and ${e_3}^T\widehat{\beta} = \dfrac{53}{80}$. We had

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 4 & 5 & 15 \\ 5 & 15 & 35 \\ 15 & 35 & 99 \end{bmatrix}$$

and $\det\left(\mathbf{X}^T\mathbf{X}\right) = 440$. Then

$$(\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{440}\begin{bmatrix} 260 & 30 & -50 \\ 30 & 171 & -65 \\ -50 & -65 & 35 \end{bmatrix}$$

With $\text{Var}(\widehat{\beta}_2) = c_{2\,2}\sigma^2 = \dfrac{35}{440}\sigma^2$.

For σ^2, we need $S^2 = \dfrac{\text{SSE}}{n-2}$.

$$\text{SSE} = y^T y - \widehat{\beta}^T X^T y = \frac{4791}{440}.$$

$$S^2 = \frac{\frac{4791}{440}}{n-2} = \frac{4791}{850} \approx 5.44432.$$

$$\text{Var}(\widehat{\beta}_2) \approx c_{2\,2}S^2 \cdots \sqrt{\text{Var}(\widehat{\beta}_2)} = 0.690201.$$

So

$$\begin{aligned} T &= \frac{e_3 \cdot \widehat{\beta} - (\widehat{\beta}_2)_0}{\sqrt{\text{Var}(\widehat{\beta}_2)}} \\ &= \frac{\frac{53}{88} - 0}{0.690201} \\ &= 0.872604\end{aligned}$$

$$\Pr\left(|T| \geq 0.872604 \mid \text{df} = 2\right) = 0.4748909.$$

This sucks! It supports the null Hypothesis.

Example: (Potency again)

x	30	30	30	50	50	50	70	70	70	90	90	90
y	38	43	29	32	26	33	19	24	23	14	19	21

I tried quadratic $y = \beta_0 + \beta_1 x + \beta_2 x^2$.

$$X = \begin{bmatrix} 1 & x & x^2 \end{bmatrix}$$

$$(X^T X)^{-1} = \frac{1}{1.92 \times 10^6} \begin{bmatrix} 1.092 \times 10^6 & -391200 & 3100 \\ -391200 & 14720 & -120 \\ 3100 & -120 & 1 \end{bmatrix}$$

Note: $\operatorname{Var}(\widehat{\beta}_2) = \dfrac{1}{1.92 \times 10^6}\sigma^2 \approx \dfrac{S^2}{1920000}$.

$$\text{SSE} = y^T y + \widehat{\beta}^T X^T y = 189.$$

$$S^2 = \frac{\text{SSE}}{n-2} = \frac{\text{SSE}}{10} = 18.9$$

$$\operatorname{Var}(\widehat{\beta}_2) = 9.84775 \times 10^{-6}.$$

$$\sqrt{\operatorname{Var}(\widehat{\beta}_2)} = 0.00313748$$

Testing

$$H_0 : \beta_2 = 0$$
$$H_a : \beta_2 \neq 0$$

I computed $\widehat{\beta}_2 = \dfrac{1}{1200}$.

$$t = \frac{\widehat{\beta}_2 - (\beta_2)_0}{\sqrt{\operatorname{Var}(\widehat{\beta}_2)}} = 0.2656$$

With df $= 10$, $\Pr(|T| > 0.2656) = 0.79$, which is HUGE again. Hence it supports the null hypothesis.

Of course, this is a math class and we love to generalive. We don't have to use a standard basis vector e_i. Let $\vec{a}$ be a constant vector $a = (a_0, \ldots, a_k)$. Then

$$a^T \widehat{\beta} = a_0 \widehat{\beta}_0 + \cdots + a_k \widehat{\beta}_k.$$

Note:

$$\begin{aligned} \operatorname{E}(a^T \widehat{\beta}) &= \sum i = 1^k a_i \operatorname{E}(\widehat{\beta}_i) && \text{(by linearity)} \\ &= \sum i = 1^k a_i \beta_i && \text{(by unbiases estimators)} \\ &= a^T \beta \end{aligned}$$

for $\beta = (\beta_0, \ldots, \beta_k)$. Also,

$$\begin{aligned} \operatorname{Var}(a^T \widehat{\beta}) &= \operatorname{Var}(a_0 \widehat{\beta}_0 + \cdots + a_k \widehat{\beta}_k) \\ &= \sum i = 1^k \sum j = 1^k \operatorname{Cov}(a_i \widehat{\beta}_i, a_j \widehat{\beta}_j) \\ &= \sum i = 1^k \sum j = 1^k a_i a_j \underbrace{\operatorname{Cov}(\widehat{\beta}_i, \widehat{\beta}_j)}_{\substack{\text{These are } c_{i\,j}\sigma^2\text{'s} \\ \text{from } (X^T X)^{-1}}} \end{aligned}$$

This entire double sum can be written as the matrix product

$$\mathrm{Var}(a^T\widehat{\beta}) = a^T(X^TX)^{-1}a \cdot \sigma^2$$

$$\implies Z = \frac{a^T\widehat{\beta} - \mathrm{E}\left(a^T\widehat{\beta}\right)}{\mathrm{Var}\left(a^T\widehat{\beta}\right)} = \frac{a^T\widehat{\beta} - a^T\beta}{\sigma\sqrt{a^T(X^TX)^{-1}a}}$$

Given $H_0 : a^T\beta = (a^T\beta)_0$ versus some alternative, use

$$T = \frac{a^T\widehat{\beta} - (a^T\beta)_0}{S\sqrt{a^T(X^TX)^{-1}a}}$$

with $n - (k+1)$ degrees of freedom.

Chapter 13

One-way Analysis of Variance

We know how to test fit the difference in 2 means

$$H_0 : \mu_1 = \mu_2 \qquad \text{i.e} \qquad \mu_1 - \mu_2 = 0.$$
$$H_a : \mu_1 \neq \mu_2 \qquad \text{i.e} \qquad \mu_1 - \mu_2 \neq 0$$

Now we endeavor to construct a test to detect difference in means over multiple groups.

Example: 24 expert typists test three new keyboard designs. Randomly assign 8 to each type of keyboard and assigned the same document to type up. The time of the task is recorded with the idea that a group with significantly less time on task would imply a better keyboard design

design	times
KB1	364 366 394 386 379 398 371 370
KB2	355 359 374 342 378 355 376 358
KB3	360 345 374 390 386 373 393 366

An obvious stat to look at is means

$$\bar{A} = 378.5 \qquad \bar{B} = 362.125 \qquad \bar{C} = 373.375$$

B "looks" better, but can we design a test to determine if it is "significantly" so? This is our goal.

The standard null hypothesis is that the true population means are the same for each group, $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k = \mu_0$. The alternative H_a is that at least one of the μ_i's are different.

Notation: Let i indicate the "group". E.g. k_i with $i = 1, 2, 3$. (k for keyboard). Let j denote the data point in that group. i.e. $Y_{1\,3} = 394$, $Y_{2\,5} = 378$.

Let n_i be the number of data points in the i^{th} group. Here $n_1 = n_2 = n_3 = 8$. Let n be the total of all data points. $n = \sum_{i=1}^{k} n_i$. Here $n = 24$.

Formally, we assume $Y_{i\,j} \sim N(\mu_i, \sigma^2)$ where each group has mean μ_i, but all groups have the same variance σ^2.
We need a stat.

To find "a" stat, we construct a likelihood ratio test. For H_0 we have our usual MLEs. If there is only μ_0

$$\widehat{\mu}_0 = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}$$

and

$$S_0{}^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \widehat{\mu}_0)^2.$$

Under the alternative hypothesis, for the individual means, we again use the standard MLE

$$\widehat{\mu}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij} = \bar{Y}_i.$$

Aside: In our motivational example, $\bar{A} = \bar{Y}_1 = 378.5$, $\bar{B} = \bar{Y}_2 = 362.125$, etc.

But we are still assuming a unique σ^2 for all groups. So the MLE for

$$S_a^2 = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \widehat{\mu}_i)^2.$$

Remark: Proving that this is an MLE is a good review and will be coming to a homework soon.

Construct a likelihood function:

$$\begin{aligned}
L(\text{All } Y_{ij} \mid \widehat{\mu}_0, S_0^2) &= \prod_{i=1}^{k}\prod_{j=1}^{n_i}\left(\frac{1}{2\pi S_0^2}\right)^{1/2}\exp\left(\frac{-(Y_{ij-\widehat{\mu}_0})^2}{2S_0^2}\right) \\
&= \left(\frac{1}{2\pi S_0^2}\right)^{n/2}\exp\left(-\frac{1}{2S_0^2}\sum_{i=1}^{k}\underbrace{\sum_{j=1}^{n_i}(Y_{ij} - \widehat{\mu}_0)^2}_{nS_0^2}\right) \\
&= \left(\frac{1}{2\pi S_0^2}\right)^{n/2}\exp\left(-\frac{nS_0^2}{2S_0^2}\right) \\
&= \left(\frac{1}{2\pi S_0^2}\right)^{n/2} e^{-n/2}
\end{aligned}$$

Under H_a, same "algebra" occurs.

$$L(\text{All } Y_{ij} \mid \widehat{\mu}_0, S_a^2) = \left(\frac{1}{2\pi S_a^2}\right)^{n/2} e^{-n/2}$$

Then the ratio of likelihood functions yields

$$\frac{L(\text{All } Y_{ij} \mid H_0)}{L(\text{All } Y_{ij} \mid H_a)} = \frac{\left(\frac{1}{2\pi S_0^2}\right)^{n/2} e^{-n/2}}{\left(\frac{1}{2\pi S_a^2}\right)^{n/2} e^{-n/2}} = \left(\frac{S_a^2}{S_0^2}\right)^{n/2} < k$$

Implies the statistic $T = \dfrac{S_a^2}{S_0^2}$ and small values of T support the alternative hypothesis (define our RR). This makes sense because if the true means are different then S_a^2 will be the correct estimator for σ^2 and S_0^2 will be larger (on average).

We have a problem... This is an entirely new stat for us.

One thing we should show is that (a)

$$S_a^2 \text{ is an estimator of } \sigma^2$$

moreover, we need it unbiased.

Note that via multiplying by the correct degrees of freedom ν, and dividing by σ^2, the numerator and denominator are χ^2 distributed. This lead to the advent of the F-distribution.

Topic: The F-distribution. Suppose $X_1 \sim \chi^2(m)$ and $X_2 \sim X^2(n)$ and are independent. Define

$$Y_i := \frac{X_1/m}{X_2/n}$$

to be an F random variable with m numerator degrees of freedom and n denominator degrees of freedom. We usually write $Y_i \sim F(m, n)$.

The derivation is akin to the derivation of the T-stat. We used the Jacobian method of transformations. (§6.6 in the text. This is usually skipped in MTH 325.)

Letting $Y_2 = X_2$, we can find the joint density function of Y_1 and Y_2, then we integrate the joint density to get the marginal density of Y_i...

$$f(y_1) = \frac{\Gamma\left(\dfrac{m+n}{n}\right)}{\Gamma\left(\dfrac{m}{2}\right)\Gamma\left(\dfrac{n}{2}\right)}\left(\frac{m}{n}\right)^{n/2}\frac{y_1^{m/2-1}}{\left(1+\left(\dfrac{my_1}{n}\right)^{(m+n)/2}\right)} \text{ and } y_i > 0.$$

Last day: Looking at at designing a test $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k = \mu_0$ versus H_a not all equal.

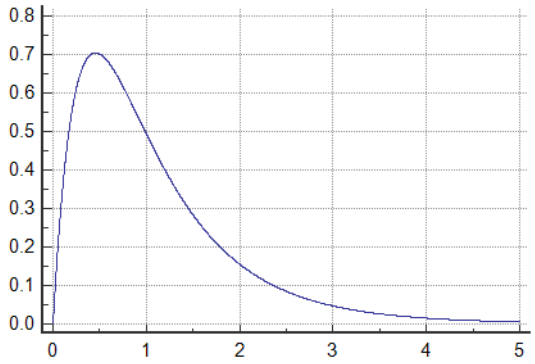
Theorem: Likelihood ration, find a new (for us) statistic that looked like a $\dfrac{\chi^2 \text{ r.v.}}{\chi^2 \text{ r.v.}}$. This lead us to the F-distribution.

$$Y \sim F(m, n) \qquad \text{when} \qquad Y = \frac{X_1/m}{X_2/n}$$

and

$$X_1 \sim \chi^2(m) \qquad \text{and} \qquad X_2 \sim \chi^2(n).$$

Graph $f(y), y > 0$:



For $\mathrm{E}\,(Y)$ and $\mathrm{Var}\,(Y)$.

$$\begin{aligned}
\mathrm{E}\,(Y) &= \mathrm{E}\left(\frac{X_1/m}{X_2/n}\right) \\
&= \mathrm{E}\left(\frac{n}{m}\cdot\frac{X_1}{X_2}\right) \\
&= \frac{n}{m}\,\mathrm{E}\left(X_1\cdot\frac{1}{X_2}\right) \\
&= \frac{n}{m}\,\mathrm{E}\,(X_1)\,\mathrm{E}\left(\frac{1}{X_2}\right) && \text{(by independence)} \\
&= \frac{n}{m}\cdot m\,\mathrm{E}\left(\frac{1}{X_2}\right) \\
&= n\,\mathrm{E}\left(\frac{1}{X_2}\right) && \text{(numerator df has no impact)} \\
&= n\cdot\underbrace{\frac{1}{n-2}}_{\text{by thm.}}
\end{aligned}$$

$$\mathrm{E}\,(Y) = \frac{n}{n-2}.$$

Also, with proof omitted,

$$\mathrm{Var}\,(Y) = \frac{2n^2(m+n-2)}{n(n-2)^2(n-4)}.$$

End F-distribution background. Then,

$$T = \frac{{S_0}^2}{{S_a}^2} = \frac{\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\widehat{\mu}_0)^2}{\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\widehat{\mu}_i)^2}$$

and the $\frac{1}{n}$ can cancel. If S_0^2 and S_a^2 were independent, then dividing by their degrees of freedom would yield an F-distribution. Sadly, they aren't. But we can algebra to

independent ratios. Start with the top,

$$\begin{aligned}
{S_0}^2 &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\widehat{\mu}_0)^2 \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left((Y_{ij}-\widehat{\mu}_i)+(\widehat{\mu}_i-\widehat{\mu}_0)\right)^2 \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left[(Y_{ij}-\widehat{\mu}_i)^2+2(Y_{ij-\widehat{\mu}_i})(\widehat{\mu}_i-\widehat{\mu}_0)+(\widehat{\mu}_i-\widehat{\mu}_0)^2\right] \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\widehat{\mu}_i)^2+2\sum_{i=1}^{k}\Big[(\widehat{\mu}_i-\widehat{\mu}_0)\quad \underbrace{(Y_{ij}-\widehat{\mu}_i)}_{\substack{=0 \text{ as it's the sum of all deviations for mean within a group from group mean}}}\quad\Big]+\sum_{i=1}^{k}n_i(\widehat{\mu}_i-\widehat{\mu}_0)^2 \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\widehat{\mu}_i)^2+\sum_{i=1}^{k}n_i(\widehat{\mu}_i-\widehat{\mu}_0)^2 \\
&= {S_a}^2+\sum_{i=1}^{k}n_i(\widehat{\mu}_i-\widehat{\mu}_0)^2
\end{aligned}$$

Then

$$\frac{{S_0}^2}{{S_a}^2} = 1+\frac{\sum_{i=1}^{k}n_i(\widehat{\mu}_i-\widehat{\mu}_0)^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\widehat{\mu}_i)^2}$$

We are pretty close to independence. Recall we have Fisher's Theorem: $\chi^2 \sim N(\mu,\sigma^2)$ then $\bar{X}$ and S^2 are independent random variables and $\bar{X} \sim N(\mu,\sigma^2/n)$.

$$(n-1)\frac{S^2}{\sigma^2} \sim \chi(1)$$

Thus the $\widehat{\mu}_i$'s up top are independent of all the variances ${S_i}^2$ in sum of the denominator. Also, $\widehat{\mu}_0$ is a function of $\widehat{\mu}_i$

$$\widehat{\mu}_0 = \frac{n_1\widehat{\mu}_1+n_2\widehat{\mu}_2+\cdots+n_k\widehat{\mu}_k}{n}$$

And the entire numerator is independent of the variance terms in the numerator when H_0 is true. Now, when H_0 is true, all the resultant variable (top and bottom) are χ^2 distributed. The last thing to do is figure out degrees of freedom. Given μ_0,

$$\frac{\widehat{\mu}_1-\mu_0}{\sigma/\sqrt{n_i}} \sim N(0,1) \qquad \text{and} \qquad \frac{n_i(\widehat{\mu}_i-\mu_0)^2}{\sigma^2} \sim \chi^2(1)$$

So, the sum of χ^2 random variables is also a χ^2 rv and the degrees of freedom add. Thus

$$\frac{1}{\sigma^2}\sum_{i=1}^{k}n_i(\widehat{\mu}_i-\mu_0)^2 \sim \chi^2(k).$$

On the other hand, to use these facts, use the say "trick" again

$$\widehat{\mu}_1 - \mu_0 = (\widehat{\mu}_1 - \widehat{\mu}_0) + (\widehat{\mu}_0 - \mu_0)$$

And

$$\sum_{i=1}^{k} n_i(\widehat{\mu}_i - \mu_0)^2 = \underbrace{\sum_{i=1}^{k} n_i(\widehat{\mu}_i - \widehat{\mu}_0)^2}_{\text{the numerator}} + n\underbrace{(\widehat{\mu}_0 - \mu_0)^2}_{\substack{\text{divide by } \sigma^2, \\ \text{this is } \chi^2(1)}}$$

In the end

$$\underbrace{\frac{1}{\sigma^2}\sum_{i=1}^{k} n_i(\widehat{\mu}_i - \mu_0)^2}_{\chi^2(k)} = \underbrace{\frac{1}{\sigma^2}\sum_{i=1}^{k} n_i(\widehat{\mu}_i - \widehat{\mu}_0)^2}_{\substack{\text{the numerator is} \\ \chi^2(k-1)}} + \underbrace{n\frac{(\widehat{\mu}_0 - \mu_0)^2}{\sigma^2}}_{\chi^2(1)}$$

Denominator is much easier,

$$\sum_{i=1}^{k} \underbrace{\sum_{j=1}^{n_i} \frac{(Y_{ij} - \mu_i)^2}{\sigma^2}}_{S_i^2 \sim \chi^2(n_i - 1)}$$

and by independences, summing all the $\chi^2(n_i - 1)$ variables yield a

$$\chi^2 \underbrace{\left(\sum_{i=1}^{k}(n_i - 1)\right)}_{\text{add all df}} = \chi^2(n-k)$$

This is a ton of work, but in the end we have

$$F = \frac{\sum_{i=1}^{k} n_i(\widehat{\mu}_i - \mu_0)^2 \Big/ k - 1}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \widehat{\mu}_i)^2 \Big/ n - k} \sim F(k-1,\, n-k)$$

when H_0 is true.

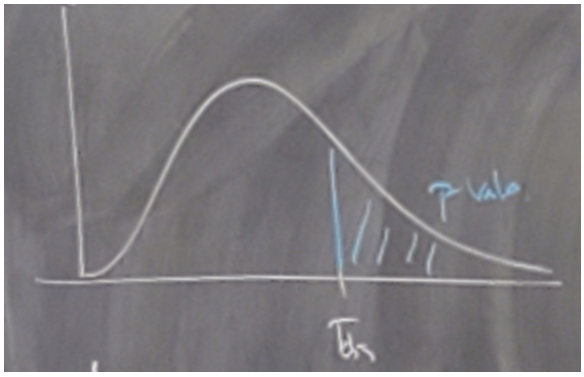
Topic: The ANOVA test.

① Compute the null and alternative estimate of all the sample means.

② Use these to compute the sum of squares in our F-statistic

③ Construct a table (R or Excel?) to compute the p-value.
F_{obs} is F observed. We computed this.

$$\Pr\left(F > F_{\text{obs}}\right),\ F - F(k-1, n-k)$$



④ Conclusion. There is standard language for all this. In the denominator,

$$\text{SSW} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \widehat{\mu}_i)^2$$

the sum of squares WITHIN a group. And the numerator

$$\text{SSA} = \sum_{i=1}^{k} n_i (\widehat{\mu}_i - \mu_0)^2$$

the sum of the squares AMONG groups. A measure of variance among the sample means. Recall originally,

$$\frac{{S_0}^2}{{S_a}^2} = \frac{{S_a}^2 + \text{SSA}}{\text{SSW}}$$

Thus we showed the factorization yield total variation.

$$S_0^2 = \text{SSW} + \text{SSA} \qquad \text{or} \qquad \text{SST} = \text{SSW} + \text{SSA}$$

ANOVA Test

$$\text{SSA} = \sum_{i=1}^{k} n_i (\widehat{\mu}_1 - \widehat{\mu}_0)^2 \qquad \text{(SS among groups)}$$

$$\text{SSW} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \widehat{\mu}_i)^2 \qquad \text{SS within groups}$$

$${S_0}^2 = \text{SST} = \text{SSA} + \text{SSW}$$

$$R^2 = \frac{\text{SSA}}{\text{SST}} \quad \text{coef. of determination}$$

Discussion: The ANOVA table

Source	SS	df	Mean Squares	F_{obs}	p-value
among	SSA	$k-1$	$\text{SSA}/(k-1)$	MSA / MSW	$\Pr(F(k-1, n-k) > F_{\text{obs}})$
within	SSW	$n-k$	$\text{SSW}/(n-k)$		
total	SST	$n-1$			

Example:

Treatment	Data		
1	2	4	3
2	6	4	
3	3	5	4

Mean squares $k = 3, n_1 = 3, n_2 = 2, n_3 = 3, n = 8, \widehat{\mu}_1 = 3, \widehat{\mu}_2 = 5, \widehat{\mu}_3 = 4.$

$$\widehat{\mu}_0 = \frac{\sum\sum Y_{ij}}{n} = \frac{n_1\widehat{\mu}_1 + n_2\widehat{\mu}_2 + n_3\widehat{\mu}_3}{n} = \frac{9 + 10 + 12}{8} = 3.875.$$

$$\begin{aligned}
\text{SSA} &= \sum_{i=1}^{k} n_i(\widehat{\mu}_1 - \widehat{\mu}_0)^2 \\
&= 3(3 - 3.875)^2 + 2(5 - 3.875)^2 + 3(4 - 3.875)^2 \\
&= 4.875. \\
\text{SSW} &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \widehat{\mu}_i)^2 \\
&= \underbrace{(2-3)^2 + (4-3)^2 + (3-3)^2}_{i=1} + \underbrace{(6-5)^2 + (4-5)^2}_{i=2} + \underbrace{(3-4)^2 + (5-4)^2 + (4-4)^2}_{i=3} \\
&= 6 \\
\text{SST} &= \text{SSW} + \text{SSA} = 10.875.
\end{aligned}$$

Note that $R^2 = \dfrac{\text{SSA}}{\text{SST}} \approx 0.326$. Make table:

Source	SS	df	Mean Squares	F_{obs}	p-value
among	4.875	2	2.4375	2.03125	0.2261023
within	6	5	1.2		
total	10.875	7			

$$\begin{aligned}
p &= \Pr\left(F(2,5) > F_{\text{obs}} = 2.03125\right) \\
&= \texttt{pf(Fobs, k-1, n-k, lower.tail=FALSE)} \qquad \text{(R code)} \\
&= 0.2261023.
\end{aligned}$$

Hence a large p-value does not support the alternative. No evidence that the means are different.

Example: Recall the typists/keyboards designs:

design	times
KB1	364 366 394 386 379 398 371 370
KB2	355 359 374 342 378 355 376 358
KB3	360 345 374 390 386 373 393 366

Then $\widehat{\mu}_1 = 378.5, \widehat{\mu}_2 = 362.125, \widehat{\mu}_3, 373.375$. $k = 3, n_1 = n_2 = n_3 = 8, n = 24$.

R-code, with little c for column. y = c (364, 355, 360, 366, 359, 345, ..., 370, 358, 366). typists = rep (1:3, 8) anova (lm (y factor(typists))). R spits out:

	df	SS	Mean Squares	F_{obs}	p-value
factor(typists)	2	532.0	266.00	1.5181	0.2422
residuals	21	3679.6	175.22		

Conclusion, again, fairly large p value. Cannot reject H_0. Practical conclusion: No evidence that our keyboard design is different in use that another. Remark: Missing from R's table is $\text{SST} = \text{SSW} + \text{SSA} = 4211.6$ and $R^2 = \text{SSA} \,/\, \text{SST} = 0.126$.

Example: Background sounds and impact on memory. 30 students are randomly divided into 3 sets of 10. Silence, classical, and jazz. They are told to "study". The "data ish"

	Quiet	Classical	Jazz
$\widehat{\mu}_i$ sample means	89.5	89.7	79.4
S_i sample standard dev.	8.91	10.25	7.06

Need $k = 3$ treatments. $n_1 = n_2 = n_3 = 10$ and $n = 30$. Then

$$\widehat{\mu}_0 = \frac{n_1\widehat{\mu}_1 + n_2\widehat{\mu}_2 + \cdots + n_k\widehat{\mu}_k}{n} = 86.2$$

$$\begin{aligned}\text{SSA} &= \sum_{i=1}^{k} n_i(\widehat{\mu}_1 - \widehat{\mu}_0)^2 \\ &= 10(89.5 - 86.2)^2 + 10(89.7 - 86.2)^2 + 10(79.4 - 86.2)^2 \\ &= 680.3\end{aligned}$$

Note we have $S_i^2 = \dfrac{\sum_j (Y_{ij} - \widehat{\mu}_i)^2}{n_i - 1}$.

$$\begin{aligned}\text{SSW} &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \widehat{\mu}_i)^2 \\ &= \sum_{j=1}^{10}(Y_{1j} - \widehat{\mu}_1)^2 + \sum_{j=1}^{10}(Y_{2j} - \widehat{\mu}_2)^2 + \sum_{j=1}^{10}(Y_{3j} - \widehat{\mu}_3)^2 \\ &= 9\left(\frac{1}{9}\sum_{j=1}^{10}(Y_{1j} - \widehat{\mu}_1)^2 + \frac{1}{9}\sum_{j=1}^{10}(Y_{2j} - \widehat{\mu}_2)^2 + \frac{1}{9}\sum_{j=1}^{10}(Y_{3j} - \widehat{\mu}_3)^2\right)\end{aligned}$$

So $\text{SSW} = 9(S_1^2 + S_2^2 + S_3^2)$. Fact, $\text{SSW} = \sum_{i=1}^{k}(n_i - 1){S_i}^2$. Here, SSW = 2108.65.

Source	SS	df	Mean Squares	F_{obs}	p-value
among	680.3	2	340.15	4.3598	0.022863
within	2108.65	27	78.09815		
total	2788.95	7			

Example: Background sounds (again)

	Quiet	Classical	Jazz
$\widehat{\mu}_i$ sample means	89.5	89.7	79.4
S_i sample standard dev.	8.91	10.25	7.06

$n_1 = n_2 = n_3 = 10$ and $n = 30$ and $k = 3$. Then

$$\widehat{\mu}_0 = \frac{\sum n_i \widehat{\mu}_i}{n} = 86.2$$

$$\text{SSA} = \sum_{i=1}^{3} n_i \left(\widehat{\mu}_1 - \widehat{\mu}_0\right)^2 = 680.3$$

$$\text{SSW} = \sum_{i=1}^{3} \sum_{j=1}^{10} (Y_{i\,j} - \widehat{\mu}_i)^2 = \sum_{i=1}^{3} 0.95_i^2 = 2108.65$$

FACT: $\text{SSW} = \sum_{i=1}^{k} (n_i - 1) {S_i}^2$. Recall:

Source	SS	df	Mean Squares	F_{obs}	p-value
among	680.3	2	340.15	4.3598	0.022863
within	2108.65	27	78.09815		
total	2788.95				

Where p-value is $\Pr\left(F(2, 27) > 4.3598\right) \approx 2.28\%$.

At $\alpha = 0.05$ level, we cannot reject H_0. Conclusion: "At least one of the group means is different than the others."

Topic: Confidence interval (§ 13.7)
We may be interested in a CI for the group mean. We clearly have an estimate for μ_i and $\widehat{\mu}_i$.

$$\text{C.I.} \equiv \widehat{\theta} \pm t^* \sqrt{\text{Var}\left(\widehat{\theta}\right)}.$$

Big question: What do we use for variance?

We have estimators ${S_0}^2, {S_a}^2$ which are MLE.
Issue: As with the original $S^1 = \frac{1}{n} \sum (X_i - \bar{X})^2$, our MLEs are biased.

H is true if H_0 is true. Both alone are "good" estimators. If H_a is true, it stands to reason that ${S_a}^2$ is "better". Using a bit of algebra, we can make an unbiased estimator of ${S_a}^2$.

Last day: we showed

$$\text{SSW} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{i\,j} - \widehat{\mu}_i)^2 = \sum_{i=1}^{k} (n_i - 1) {S_i}^2.$$

This is a weighted sum of unbiased estimators ${S_i}^2$ for σ^2. As individually, each ${S_i}^2 \sim \sigma^2$ (unbiased). Then the collective average should be an even better approximation.

$$\sigma^2 = \frac{\sum_{i=1}^{k}(n_i - 1){S_i}^2}{(n_1 - 1) + \cdots + (n_k - 1)} = \frac{\sum_{i=1}^{k}(n_i - 1){S_i}^2}{n - k} = \text{MSW}$$

FACT: We can use the MSW as an unbiased estimator for σ^2.
Aside: Isn't this just the pooled estimator written largely?

$${S_p}^2 = \frac{(n_1 - 1){S_1}^2 + (n_2 - 1){S_2}^2}{n_1 + n_2 + k}$$

Here

$$S_{\cdot}^2 = \frac{(n_1 - 1){S_1}^2 + \cdots + (n_k - 1){S_k}^2}{n_1 + n_2 + \cdots + n_k - k}$$

We are constructing hypothesis of different of CI. The "t" inherits the df from S^2. For our CI we need $\text{df} = n - k$.

Example: Background sounds (again)

a.) Find a 95% CI for the mean score for someone listening to Jazz.

Here, $n_3 = 10$ and $t_{\alpha/2}(\text{df}) = t_{0.025}(27) \approx 2.052$.
Then, $\widehat{\mu}_3 \pm t_{0.025}(27)\sqrt{S^2/n_3}$ with S^2/n being the total variance affiliated with samling dist.

We have $79.4 \pm 2.052\sqrt{\dfrac{78.09815}{10}}$ which is

$$79.4 \pm 5.734 \qquad \text{or} \qquad (73.665,\ 85.134)$$

Note $\widehat{\mu}_1 = 89.5$ and $\widehat{\mu}_2 = 89.7$ this is still a couple standard errors away.

b.) Is the proof of the other 2 μ_i's are different? Note: We could ANOVA again without Jazz, or do chapter 9 stuff:

Doing Ch. 9 stuff, consider $\widehat{\mu}_1 - \widehat{\mu}_2 = -0.2$. Then to make a CI we use the exact same formula from before (Ch9) but we use the "ANOVA" estimator for $S^2 = \text{MSE}$ with $\text{df} = n - k$.

$$\begin{aligned}
\widehat{\mu}_1 - \widehat{\mu}_2 \pm t_{0.025}(27)\sqrt{\frac{S^2}{n_1} + \frac{S^2}{n_2}} &= \widehat{\mu}_1 - \widehat{\mu}_2 \pm t_{0.025}(27)S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \\
&= -0.2 \pm 20.052\sqrt{\frac{75.09815}{5}} \\
&= -0.2 \pm 7.95 \\
&\equiv (-8.15,\ 7.75)
\end{aligned}$$

And notice that $0 \in$ CI.

Summary: 2-sided $1-\alpha$ level CI.

$$\widehat{\mu}_i \pm t_{\alpha/2}(\mathrm{df})\frac{S}{\sqrt{n_i}} = \widehat{\mu}_i - \widehat{\mu}_j \pm t_{\alpha/2}(\mathrm{df})S\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}$$

Where $S^2 = \mathrm{MSW} = \dfrac{\mathrm{SSW}}{n-k}$ with $\mathrm{df} = n-k$.

Example: Arbitrary numbers (no story)

A	80	85	71	64
B	70	72	75	
C	83	70		

Then $n_1 = 4$, $n_2 = 3$, $n_3 = 2$, $n = 9$, $\bar{A} = 75$, $\bar{B} = 217/3$, $\bar{C} = 153/2$ and

$$\widehat{\mu}_0 = \frac{4\bar{A}+3\bar{B}+2\bar{C}}{9} = \frac{690}{9}.$$

	SS	df	MS	F_{obs}
among	SSA = 23.0556	2	11.5278	0.192575
within	SSW = 359.167	6	39.8611	

The above was computed on mathematica. Using R for the p-value, $p \approx 82.97\%$. We cannot reject H_0.

The R code:

```
scores = c(80, 85, 71, 64, 70, 72, 75, 83, 70)
treat = c(rep("A", 4), rep("B", 3), rep("C", 2))
table = data.frame(saved, treat)
results = aov(scores ~ treat, data = table)
summary(results)
```

The output is

	df	Sum Square	Mean Square	F	$\Pr(>F)$
treat	2	23.1	11.53	0.193	0.83
residual	6	359.2	59.86		