Stat 135: Notes

Avi Garg

April 15, 2020

1 Lecture: January 22nd, 2020

1.1 Introduction: Parameter Estimation and Std Error

A coin lands heads with probability p, and is tossed 100 times and lands heads 45 times. What can you say about p?
Answer: We can estimate p with $\hat{p}$ and fin a SE for $\hat{p}$.
$\hat{p} = \bar{X} = \frac{1}{100}\sum_{i=1}^{100} x_i = \frac{45}{100} = .45$ $\hat{p}$ is a random variable and it has a distribution called the sampling distribution. Since $\hat{p}$ is summation of multiple independent variables, we know it will follow an approximately normal distribution.
$\text{Var}(\hat{p} = Var(\frac{1}{100}\sum_{i=1}^{100} x_i = \frac{1}{10000}Var(\sum_{i=1}^{100} x_i) = \frac{1}{100}Var(X_i)$. Since we know that each coin flip is a Bernoulli variable, we know that the variance is $p(1-p)$ which means that we have the approximate $SE(\hat{p}) = \sqrt{\frac{(1-p)p}{100}}$. However, this relies on the population parameter so we can rewrite this in a different way:
$SE(\hat{p}) = \sqrt{\frac{(\hat{p}(1-\hat{p})}{100}} = \sqrt{\frac{(.45)(1-.45)}{100}} = .0497$
Conservative Estimate of SE($\hat{p}$):
$Var(\hat{p}) = \frac{p(1-p)}{100}$. We know that this takes the shape of a parabola with intercepts at 0 and 1 and a maximum at .5. We know that the maximum variance we can ever have is .05 because the maximum variance is .25, so $\sqrt{\frac{.25}{100}} = .05$

1.2 Summary

We have a dichotomous case, box 0, 1
p is equal to the proportion of 1's in the box
draw some sample size n, x_1, x_2, x_3.. w/ replacement from box with N numbers $\hat{p} = \bar{x}$ bootstrap SE($\hat{p}$) = $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
The Conservative estimate is $\frac{.5}{\sqrt{n}}$. However, this is only concerning w/ replacement. The minute that we have no replacement, we need to add a correction factor to it (generally referred to as a SRS).
$x_1, x_2, ..x_n$ are dependent now, and we add the correction factor: $Var(\bar{X}) = \frac{p(1-p)}{n}[\frac{N-n}{N-1}]$

Standard Error is equal to standard deviation, but standard deviation is concerning a population parameter instead of a sampling of a population. How far a sample parameter estimate is from the true population estimate.

1.3 Examples

Assume that we have a box with 5 numbers, 00001 and that we are sapling 2 w/o replacement. We have to write out the sampling distribution.
$(0,0) - - > \frac{6}{10}, \hat{p} = 0$
$(0,1) - - > \frac{4}{10}\hat{p} = .5$
$E(\hat{p}) = 0 * \frac{6}{10} + .5 * \frac{4}{10} = \frac{1}{5}$ $Var(\hat{p}) = E(\hat{p}^2) - (E(\hat{p})^2 = 0 * \frac{6}{10} + (\frac{1}{2})^2 \frac{4}{10} - (\frac{1}{5})^2 = .06$

2 Lecture: January 24th, 2020

2.1 Sec 7.3.3

Let x_1, x_2... be a SRS fro m distribution with mean μ and variance σ^2. We can show that $E(\bar{X}) = \mu$ and $Var(\bar{X}) = \frac{\sigma^2}{n}\frac{N-n}{N-1}$

2.1.1 Example

Population is 393 hospitals, and let x_i be the number of patients discharged from i^{th} hospital.
$\mu = 814.6$
$\sigma = 590$

A SRS of n = 50 is taken, we want to find the probability of:
$P(|\bar{X} - \mu| > 100)$. $\hat{\mu} = \bar{x}$ and $\sigma_{\bar{X}} = \sqrt{\frac{\sigma^2}{n}\frac{N-n}{N-1}} = 77.95$
We know that the curve is approximately normal and as a result we can calculate the z-score and then use inverse Gaussian to achieve an answer. It ends up being 1.25 standard deviations above, which means that the probability of that event occuring is .1 and the event it is less than this is symmetrically less, thus the total probability is .2

2.1.2 Example: Dichotomous Case

let p is the proportion of hospitals with fewer than 1000 discharges. We know that $p = .65$

$$P(|\hat{P} - p| > .13)$$

This means that we $E(\hat{P}) = .65$ and that $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}\frac{N-n}{N-1}} = \sqrt{\frac{.65*.25}{50}\frac{343}{392}} =$ .063. Thus we can find the z-scores which gets us 2.06 which results in a total p-value of .039
The equivalence of SE of a parameter is to give the confidence interval
Confidence Intervals:
A CI for the mean μ or p is a random interval, calculated from the sample that contains μ with a specified probability. $For 0 \leq \alpha \leq 1$ let $Z(\alpha)$ be the number such that the area under the standard normal curve to the right of $Z(\alpha) = \alpha$

3 Lecture: January 29th, 2020

3.0.1 Method of Moment (MOM) estimators and consistency of mom estimators (8.4)

8.4: The method of moments (mom)
Review of Gamma Distribution: X ~ Gamma(r, λ), and the X models the r^{th} arrival in a $Pois(\lambda)$ process. Thus we have $f(X) = \frac{\lambda^r}{\Gamma(r)}x^{r-1}e^{-\lambda x}$.
$\Gamma(r) = (r-1)!$
$E(X) = \frac{r}{\lambda}$
$Var(X) = \frac{r}{\lambda^2}$

Method of Moment:
If we want to estimate l parameters, $(\Theta_1, Theta_2...\Theta_l$ of a prob distribution. $f(X|\Theta_1, \Theta_2...\Theta_l)$ from i.i.d sample $X_1, X_2..X_n$ from this distribution.
Step 1: Compute the first l moments:
$M_X = E(x^k) \forall k \in (1, l)$
$M_1 = g_1(\Theta_1, \Theta_2..\Theta_l)$
$M_2 = g_2(\Theta_1, \Theta_2..\Theta_l)$
....
$M_l = g_l(\Theta_1, \Theta_2..\Theta_l)$
Examples:

1. $X\ Pois(lambda)$
   $M_1 = E(X) = \lambda$

2. $X\ Gamma(r, \lambda) l = 2$
   $M_1 = E(X) = \frac{r}{\lambda}$
   $M_2 = E[x^2] = \frac{r+r^2}{\lambda^2}$

   Step 2: Use Algebra to compute the system of equations
   $\Theta_1 = h_1(\mu_1, \mu_2...\mu_l)$
$\Theta_2 = h_2(\mu_1, \mu_2...\mu_l)$
...
$\Theta_l = h_l(\mu_1, \mu_2...\mu_l)$

   Examples

1. $\mu_1 = \lambda$

2. $\mu_1 = \frac{r}{\lambda}$
   $\mu_2 = \frac{r+r^2}{\lambda^2}$
   $\mu_2 = \frac{mu_1}{\lambda} + \mu_1^2$
   $\lambda = \frac{\mu_1}{\mu_2 - \mu_1^2}$
   $r = \frac{\mu_1^2}{\mu_2 - \mu_1^2}$

   Step 3: Insert into the estimator for the moments $\mu_1...\mu_k$.

$\hat{\mu_1} = \frac{1}{n}\sum_{i=1}^{n} x_i$
$\hat{\mu_2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$
...
$\hat{\mu_l} = \frac{1}{n}\sum_{i=1}^{n} x_i^l$
$\hat{\Theta}_1 = h_1(\hat{\mu_1}, \hat{\mu_2}..\hat{\mu_l})$
$\hat{\Theta}_2 = h_2(\hat{\mu_1}, \hat{\mu_2}..\hat{\mu_l})$
...
$\hat{\Theta}_l = h_l(\hat{\mu_1}, \hat{\mu_2}..\hat{\mu_l})$

Mom estimators have nice properties:

- They converge to the parameter in probability
  This means that for $\epsilon > 0, P(|X_n - X| > \epsilon) \to 0, n \to \infty)$

Ex
The sample mean converges to the population mean in probability.
Let $X_1, X_2...X_n$ be i.id with mean μ and var σ^2). We want to prove that:
$X_{(n)} \to E(X)$
$P(|X_n - X| > \epsilon) \to 0, n \to \infty)$
$P((\bar{X}_{(n)} - \mu)^2 \geq \epsilon^2)$. Thus w have, by Markov's inequality, that this must be less than or equal to $\frac{E((\bar{X}_{(n)} - \mu)^2}{\epsilon^2} = \frac{Var(\bar{X}_{(n)}}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}$ As $n \to \infty$ this expression goes towards 0.

4 Lecture: January 31, 2019

4.1 Sec 8.4: Proof MOM estimators are are consistent

Def An estimator $\hat{\Theta}$ of a parameter Θ is consistent if $\hat{\Theta} \to \Theta$
We argue that mom estimators are consistent.
Theorem If $X_n \to^p X$ and h is continuous then $h(x_n) \to^p h(x)$
Theorem Generalized Weak Law of Large Numbers
$\frac{1}{n}\sum_{k=1}^{n} X_i^k \to^p E(X^k)$. This means that the sample mean will converge in probability to the k^{th} moment.
So we have $(\hat{\mu_1}, \hat{\mu_2}, ...\hat{\mu_l}) \to^p (\mu_1, \mu_2...\mu_l)$
Let h be continuous then:
$h(\hat{\mu_1}, \hat{\mu_2}, ...\hat{\mu_l}) \to^p h(\mu_1, \mu_2...\mu_l)$
The first of these is the mom estimator ($\hat{\Theta}_{mom}$ and the second is Θ.

4.2 Sec 8.4: Example of mom calculation

Example: No. 169 chap 8
Gamma(r, λ), $f(x) = f(x|r, \lambda) = \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x}$

$$\int_0^\infty f(x)dx = 1 \rightarrow \int_0^\infty \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x} dx = 1$$

$$\int_0^\infty x^{r-1} e^{-\lambda x} dx = \frac{\Gamma(r)}{\lambda^r}$$

Problem
Consider an i.i.d sample of RV's w density $f(x|\sigma) = \frac{1}{2\sigma} e^{-|x|} \sigma, \sigma > 0$ and we want to find $\hat{\sigma}$.

Step 1: Find the moment, E(X)
$\mu_1 = E(X) = \int_{-\infty}^{\infty} \frac{x}{2\sigma} e^{\frac{-|x|}{\sigma}} dx = 0$ because we have an odd function. However, we want to represent our estimator as function of X, so this is essentially useless. We can go to the next moment.
$\mu_2 = E(X^2) = \int_{-\infty}^{\infty} \frac{x^2}{2\sigma} e^{\frac{-|x|}{\sigma}} dx$ Because this is even we can see that this is equal to :
$\int_0^\infty \frac{x^2}{\sigma} e^{\frac{-x}{\sigma}} dx = \frac{\Gamma(3)}{\sigma^4} = 2\sigma^2$ since this is a Gamma distribution with $\lambda = \frac{1}{\sigma}, r = 3$
Step 2 $\sigma = \sqrt{\frac{\mu_2}{2}}$
Step 3 $\hat{\sigma} = \sqrt{\frac{\frac{1}{n}\sum_{i=1}^n X_i^2}{2}}$

4.3 Non parametric Bootstrap for 95% confidence interval

To find SE in a confidence interval, we often need to bootstrap.
Step 1:
Take a sample $X_1, X_2..X_n$ of size n one time from your population. Calculate $\hat{\Theta}$.
Step 2:
Resample from your sample many many times with replacement and compute the sample estimate of your parameter(Θ), $\Theta_1^*, \Theta_1^*...\Theta_B^*$ where B is the number of times we resample.

Step 3:
Subtract $\hat{\Theta}$ from each f our Θ_i^* and we get B numbers. This will have some type of distribution and we can then create a confidence interval from this.
Step 4: Find 2.5 and 97.5 biggest value and call this a and b.
Step 5
Given a, b, the 95% CI of Θ is $(\hat{\Theta} - b, \hat{\Theta} - a)$ which is equivalent to P(a ¡ $\hat{\Theta} - \Theta < b) = 95\%$.

5 Lecture: February 3rd 2020

5.0.1 Section 10.2: The Empirical CDF (ecdf)

When we do non parametric bootstrap, why does it yield us an equivalent answer? How does resampling from a sample equal sampling from a population?
Suppose our population cdf, F, is a non decreasing function:

Let $X_1, X_2...X_n$ be our original sample: $(X_1, X_2...X_n\ F)$. Define ecdf as $F_n(x) = \frac{1}{n}(\#X_i \leq x)$ (i.e the fraction of data points that are less than x)
Facts(P380):
1. F_n is an unbiased estimator of F.
2. $SE(F_n) \to 0$ as $n \to \infty$

When doing non parametric bootstrap instead of resampling from F given we resample from F_n size n and that works for resampling of size n.
The CDF is unique to a distribution and represents the population, and the ecdf represents the sample. We know that these two converge under enough simulation, which is why bootstrapping works.

5.1 Computing $SE(\hat{\Theta})$ by hand

Ex: An i.i.d sample $x_1 = 4, x_2 = 7, x_3 = 4, x_4 = 2, x_5 = 3$ is taken and follows a Poisson (λ) distribution. Find a mom estimator of λ and approximate the $SE(\lambda)$
Facts:
$\hat{\lambda} = \bar{X}$
$Var(X) = \lambda$

$\sum_{i=1}^{5} x_i = 20$. Lambda is equal to 4 because it is the average, giving us $\frac{20}{4}$.
$SE(\hat{\lambda})$ is equal to $\sqrt{Var(\bar{X})} = \sqrt{\frac{Var(X)}{n}} = \sqrt{\frac{\lambda}{n}} \approx \sqrt{\frac{\hat{\lambda}}{n}} = \sqrt{\frac{4}{5}}$
We can also substitute the sample variance for Var(X). $s^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2 = \frac{14}{4}$ and this is relatively close to 4.
Ex
Suppose that X is a discrete RV with:
$P(X=0) = \frac{2}{3}\Theta$
$P(X=1) = \frac{1}{3}\Theta$
$P(X=2) = \frac{2}{3}(1-\Theta)$
$P(X=3) = \frac{1}{3}(1-\Theta)$
The following observations are taken:
3, 0, 2, 1, 3, 2, 1, 0, 2, 1
Find the MOM estimator of Θ and approximate $SE(\hat{\Theta})$) $E(X) = \frac{1}{3}\Theta + \frac{4}{3} - \frac{2}{3}\Theta + 1 - \Theta = -\frac{4}{3}\Theta + \frac{7}{3}$
$\Theta = \frac{1}{2}(\frac{7}{3} - \mu_1) = \frac{5}{12}, \mu_1 = \frac{3}{2}$

$Var(\hat{\Theta}) = Var(\frac{1}{2}(\frac{7}{3} - \bar{X})) = \frac{1}{4}Var(\bar{X})$

2 options to compute Var(X)
1. Approximate $Var(X)$ by s^2 and get $SE(\hat{\theta}) = .171$ = Done in R $sqrt(\frac{1}{4\Theta}Var(c(3,0,2,1,3,2,1,0,2$
2. Find $Var(X) = E(X^2) - E(X^2)$ and get $SE(\hat{\Theta}) = .173$

6 Lecture: February 5th, 2020

6.1 Sec 4.6: Delta Method to find $SE(g(\bar{X}$

Ex $X_1, X_2, , X_3...$ are i.i.id $Exp(\lambda)$
Fact: $\mu = E(X) = \frac{1}{\lambda}$
Find $\hat{\lambda}$ and SE(λ).
$\mu_1 = \frac{1}{\lambda} \to \lambda = \frac{1}{\mu_1} \to \hat{\lambda} = \frac{1}{\bar{X}}$
$(SE(\hat{\lambda}))^2 = Var(\frac{1}{\bar{X}})$. However, we need the delta method to approximate this.

Theorem: Delta Method
$X_1, X_2...X_n$ i.i.d mean μ, SD = σ. g is smooth around μ and $g'(\mu) \neq 0$:
$var(g(\bar{x})) \approx (g'(\mu))^2 \frac{\sigma^2}{n}$

The proof is essentially mirroring a taylor series fro $g(\bar{x})$ around μ.

$g(\bar{x}) = g(\mu) + g'(\mu)(\bar{x} - \mu) + \frac{g''(\mu(\bar{x}-\mu)^2}{2!}$
For large n, $\bar{X}$ is close to μ so $g(\bar{x}) \approx g(\mu) + g'(\mu)(\bar{x} - \mu)$
$var(g(\bar{x})) \approx var(g(\mu) + g'(\mu)(\bar{x} - \mu))$
$var(g(\bar{x})) = g'(\mu)^2 var(\bar{X} - mu)$
$Var(\bar{X} - mu)$ is equal to the variance of $\bar{x}$ because μ is just a constant, and thus it equal to $\frac{\sigma^2}{n}$.

Back to $X\ Exp(\lambda)$ example"
We saw $\hat{\lambda} = \frac{1}{\bar{X}}$ so let $g(\bar{X}) = \frac{1}{\bar{X}}$ in delta method.
$Var(g(\bar{X})) \approx (g'(\mu))^2 \frac{\sigma^2}{n}$
$g'(x) = -\frac{1}{x^2} \to (g'(\mu))^2 = (-\frac{1}{\mu^2})^2 = \lambda^4$
$Var(\hat{\lambda}) \approx \lambda^4 \cdot \frac{\sigma^2}{n} = \frac{\lambda^2}{n}$
$Var(\hat{\lambda}) = \frac{\frac{1}{\bar{X}}^2}{n}$, $SE(\hat{\lambda}) = \frac{1}{\bar{X}\sqrt{n}}$

Ex Let $X_1, X_2..X_n$ i.i.d R.V's w/ density $f(X|\Theta) = (\theta+1)X^{\Theta}$, $0 \leq X \leq 1$
Find $\hat{\Theta}$ an use δ -method to approximate at $SE(\hat{\Theta})$
$E(X) = \mu_1 = \int_0^1 (\Theta + 1)x^{\Theta+1}dx = \frac{\Theta+1}{\Theta+2}$
$E(X^2) = \mu_2 = \frac{\Theta+1}{\Theta+3}$
$\Theta = \frac{1-2mu_1}{\mu_1-1} \to \hat{\Theta} = \frac{1-2\bar{X}}{\bar{X}-1}$
$SE(\hat{\Theta})$
$Var(\hat{\Theta}) - (g'(\mu))^2 \frac{\sigma^2}{n}$
$Var(\hat{\Theta}) - (\frac{1}{(\mu-1)^2}) \frac{\frac{\Theta+1}{\Theta+3} - (\frac{\Theta+1}{\Theta+2})^2}{n}$

6.2 Sec 8.5: Maximum Likelihood Estimator (MLE)

Suppose R.V $X_1, X_2..X_n$ have joint density $f(X_1, X_2, X_n|\Theta)$
Given observed data:
$X_1 = x_1, X_2 = x_2, ...X_n = x_n$, fixed numbers
$lik(\Theta) - f(x_1...x_n|\Theta)$ This is a function of Θ for fixed $x_1, ...x_n$ and the max value of $lik(\Theta)$ represents the value of Θ that maximizes the likelihood of observing your data.
$If X_i$ are i.i.id
$lik(\Theta) = \Pi_{i=1}^n f(X_i|\Theta)$.
Take the log of both sides and take the derivative and find the Θ that max-

imizes the likelihood of Θ
$log(lik(\Theta)) = \sum_{i=1}^{n} log(f(x_i|\Theta)$

7 Lecture: February 7th, 2020

Ex: # 16 b
Consider an i.i.d sample RV with densities
$f(X|\sigma) = \frac{1}{2\sigma} e^{\frac{-|X|}{\sigma}}$
$Find \hat{\sigma}_{ML}$
$lik(\sigma) - \Pi_{i=1}^{n} f(x_i|\sigma) = \frac{1}{2\sigma}^n e^{\frac{1}{\sigma}(|X_1||X_2|..|X_n|}$
$l(\sigma) = nlog(\frac{1}{2\sigma}) - \frac{1}{\sigma}\sum_{i=1}^{n}|X_i|$
$l'(\sigma) = -\frac{n}{\sigma} + \frac{1}{\sigma^2}\sum_{i=1}^{n}|X_i| = 0$
$\sigma_{ml} = \frac{\sum_{i=1}^{n}|X_i|}{n}$

7.1 Property of MLE

- Consistency of MLE:
  We say an estimate $\hat{\Theta}$ is consistent if over time it will equal the true Θ
  Proof?
  Let $X_1, X_2..X_n$ are i.i.d F_{Θ_0} where Θ_0 is unknown, assume $L(\Theta)$ is smooth and behaves in a nice way.
  $\hat{\Theta}$ maximizes $\frac{1}{n}\sum_{i=1}^{n} log(f(x_i|\Theta))$ and is dependent on the data, and is thus a random variable. However, the expectation of that is:
  $l(\Theta) = E_{\Theta_0}\frac{1}{n}\sum_{i=1}^{n} log(f(x_i|\Theta)) = E_{\Theta_0}(log(f(x|\Theta)))$
  By the weak law of large numbers, $l_n(\Theta) = l(\Theta)$
  Fact: $l(\Theta) < l(\Theta_0) \forall \Theta$

8 Lecture February 10th, 2020

8.1 Equivalence for MOM and MLE

$g(\hat{\Theta}) = g(\hat{\Theta})$

8.2 Sec 8.5: CI of parameter $\mathbf{N}(\mu, \sigma^2)$

(a) Chi Square Distribution

$$\begin{aligned} Z &\sim N(0,1) \\ Z^2 &\sim \chi_1^2 \\ Z^2 &\sim Gamma(\frac{1}{2}, \frac{1}{2}) \\ Z_1^2 + Z_2^2 + ...Z_n^2 &\sim \chi_n^2 \\ \chi_n^2 &\sim Gamma(\frac{n}{2}, \frac{1}{2}) \end{aligned}$$

(b) Important Identity
If $X_1, X_2..X_n \to\sim N(\mu, \sigma^2)$ then
$\frac{n-1}{\sigma^2}s^2 \sim \chi_{n-1}^2$ and S^2 has n-1 degrees of freedom

(c) The distribution of the μ in random samples of a normal distribution follow a t-distributon where $t_{n-1} = \frac{z}{\sqrt{\frac{u}{n-1}}}$ where u is a χ_{n-1}^2 . We know that μ follows a t distribution of this sort.

9 Lecture: February 12th, 2020

95% CI in R is given by:
$2\hat{\lambda} - quantile(\lambda^*, c(.975, .025))$ and λ^* is a thousand or more resampling while $\hat{\lambda}$ is our original sample estimate.

9.1 Large Sample Theory for MLE

Theorem 1 *The MLE $\hat{\theta}$ for Θ is asymptotically (for large n) unbiased and normal. More precisely $\hat{\Theta} \approx N(\Theta_0, \frac{1}{nI(\Theta_0)})$ where $I(\Theta_0)$ is the Fischer Info (FI) at true the value Θ_0*
Fischer Info:
Fi is the measure of the amunt of info a single observation x has to estimate $\hat{\Theta}_{ML}$

ex Lets look at $log(f(x|\Theta))$ for two circumstances where x is really informative and uninformative

informative

$$X \sim N(\mu, 1)$$
$$l(\mu) = log(f(x|\mu)) = log(\frac{1}{\sqrt{2\pi}} e^{(x-\mu)^2})$$
$$= log(\frac{1}{\sqrt{2\pi}} - (x-\mu)^2$$

uninformative
$X \sim N(\mu, 25)$
$l(\mu) = log(\frac{1}{\sqrt{50\pi}} - (\frac{x-\mu)}{25})^2$

Each curve is providing its own vote for the true parameter location (i.e the peak). How can we measure how tight the peak is? We have to look at the slopes of the curves (called the score function) and the $score_x(\Theta) = \frac{d}{d\Theta} log(f(x|\Theta)) = l'(\Theta)$
We are interested in Θ and $E(l'(\Theta)) = 0$

10 Lecture: February 14th, 2020

$I(\theta) Var(l'(\theta))$
$I(\theta) = -E(l''(\theta))$

10.1 Cramer-Rao Inequality (CR) Inequality

Main Points:
Let $X_1, ...X_n$ iid $f(x, \theta)$

1. $\hat{\theta}_{ML} \approx N(\theta, \frac{1}{nI(\theta)})$ for large n

2. Cramer Rao inequality
   Let $\hat{\theta}_{ML}$ be unbiased then $Var(\hat{\theta}_{ML}) \leq \frac{1}{nI(\theta)}$

Definition: An unbased estimator whose variance achieves the CR lower bound is called efficient
From 1, above we see that $\hat{\theta}_{ML}$ is asymptotically efficient.

11 Lecture: February 19th, 2020

11.1 Sec 8.7: Mean Square Error (MSE) of an Estimator

The MSE is used to measure how good of an estimator $\hat{\theta}$ is.
Definition: $MSE(\hat{\theta}) = E((\hat{\theta} - \theta)^2) = Var(\hat{\theta}) + Bias(\hat{\theta})^2$
$Bias(\hat{\theta}) = E(\hat{\theta}) - \theta$

$MSE(\hat{\theta}) = E((\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta)^2)$
Ex:
$X_1, X_2 ... X_n$ $Bern(\theta)$

$\hat{\theta} = \bar{X}$
$\tilde{\theta} = X_1$
$E(\hat{\theta}) = \theta$
$E(\tilde{\theta}) = \theta$
$Var(\hat{\theta}) = \frac{\theta(1-\theta)}{n} Var(\tilde{\theta}) = \theta(1-\theta)$
Thus $\hat{\theta}$ has the smallest MSE (both are unbiased thus we don't have to include that part into our calculation
Conclusion: Among all unbiased estimators the MLE for large sample size has the small possible MSE since it is efficient. (it achieves CR lower bound). However, it is to note that the Cr inequality only applies to unbiased estimators so if you allow the estimator to be biased you might get an even smaller MSE than you might get among unbiased estimator.

11.2 Sec 8.8 Sufficiency

There are two primary motivations for sufficient estimators

(a) Once you collect you data you can form a sufficient statistic and then throw away your data; you only need the sufficient statistic to estimate

θ

(b) (Rao-Blackwell Theorem)
You can make an estimator $\hat{\theta}$ better by taking the conditional expectation given a sufficient statistic T
$\tilde{\theta}(T) = E(\hat{\theta}(X_1..X_n)|T)$

Definition A statistic $T(X_1...X_n)$ is a function of your data only (e.g $\bar{X}$, $min(X_1...X_n)$, X_3, ...) but it cannot be like $\bar{X} - \mu$ because this involves information we don't have within our data.

Definition T is a sufficient statistic for θ if the conditional distribution of $X_1...X_n|T$ doesn't depend on θ
ex: X $N(\theta, 1)$
$T = X$
$X|T = x$
Ex $X_1 = \#$ red cars crossing golden gate bridge in 1 min Pois(θ)
$X_2 = \#$ blue cars crossing golden gate bridge in 1 min Pois(θ)
X_1, X_2 are iid
$T = X_1, X_2 \quad Pois(2\theta)$
$X_1|T = 4 \quad$ Bin(4, $\frac{1}{2}$) This does not depend on θ which means that it is a sufficient statistic of X_1

Recall Baye's Rule:
$P(A|B) = P(A \cap B)/P(B)$
and for densities it says:
$f(x_1..x_n|T = t, \theta) = \frac{f(X_1=x_1, X_n=x_n, T=t|\theta)}{f(T=t|\theta)}$

12 Lecture: February 21st, 2020

12.1 Sec 8.8 Factorization Theorem

$X^n = (x_1, x_2..x_n)$
$Y^n = (y_1, y_2...y_n)$

Theorem 2 *T is sufficient if*

$$f(X^n|\theta) = g(T(X^n), \theta)h(x^n) \tag{1}$$

Thus it factors into two functions, one that doesn't depend on θ and the other that does but only through $T(x)$

Corollary A:
If T is sufficient for θ the MLE estimate is a function of T.

13 Lecture: February 24th, 2020

13.1 Sec 8.8

T is MSS if SP = LP;
So T is MSS iff the following is true:
$T(X^n) = T(y^n) iff \frac{f(y^n|\theta)}{f(x^n|\theta)}$ doesn't depend on θ.
X_1, X_2 $Bern(\theta)$
Find MSS:
$\frac{f(y_1,y_2|\theta)}{f(x_1,x_2|\theta)} = \frac{\theta^{\sum y_i}(1-\theta)^{2-\sum(y_i}}{\theta^{\sum x_i}(1-\theta)^{2-\sum(x_i}}$
This doesn't depend on θ when $\sum y_i = \sum x_i$
1

Rao-Blackwell teorem describes how to improve an estimator using a sufficient statitics $\tilde{\theta} = E(\hat{\theta}|T)$ has smaller MSE than $\hat{\theta}$

14 Lecture: February 26th, 2020

14.1 Hypothesis Testing

Sec 9.2 Frequentist approach
Definition: A hypothesis is a statement about a population parameter. The goal is to decide based on your sample $X_1..X_n$ which of two complementary hypothesis is true.
H_0 = Null Hypothesis
H_1 = Alternate Hypothesis
Decision Functions
Hypothesis testing makes a binary decision, either accept or reject the null

hypothesis

$$d(x) = \begin{cases} 1 & H_0 rejected \\ 0 & H_0 accepted \end{cases}$$

be the decision function
It is always possible to make an error
$\alpha = P_0(d(x) = 1)$ a Type 1, false positive (Null was true but we rejected it)
$\beta = P_1(d(x) = 0)$ a Type 2, false negative (Null wasn't true bu we accepted it)

α, β are inversely correlated, meaning we can't lower both. Thus we fix α at a certain level and design test to minimize β as much as we can.
We call α a significance level

14.2 Likelihood Ratio Test

If we have $f_0(x)$ which we assume is the null distribution and $f_1(x)$ which is the alternate distribution, we can calculate $\wedge = \frac{f_)(x)}{f_1(x)}$
$P_0(\lambda \leq c)$ reject lambda is small
We can look at this like, what is the probability that we get the value of X as x_1 given both the null and alternate distribution. This means that we can set an α to view this probability and choose a c that agrees in a theoretical sense.
Values for Λ in whih we reject H_0 is called the rejection region.

14.3 Power of Test

Power $= 1 - \beta = 1$ - $P_1(d(x) = 0) = P_1(d(x) = 1)$
Definition
Suppose under the null $X \sim Unif[0, 1]$ and under the alternative X has density $f(x) = 2x, 0 < x < 1$

1. What is LRT at $\alpha = .1$ level of significance
2. What is the power of the test?

LRT reject H_0 if $\Lambda = \frac{f_0(x)}{f_1(x)} \leq c = \frac{1}{2x}$
$.1 = P(\frac{1}{2x} < c)$ c $= \frac{1}{1.8}$

15 Lecture: March 13th, 2020

15.1 Sec 11.1, Sec 11.2

2 sample z-test if we know the variance.
$X_1, X_2...X_n \sim N(\mu_x, \sigma^2)$
$Y_1, Y_2...Y_m \sum N(\mu_y, \sigma^2)$
These are both independent samples that we drew, with one being the control and the other being the treatment group. We want to see if there is a statistically significant difference between them
$\bar{X} - \bar{Y} \sim N(\mu_x - \mu_y, \sigma^2(\frac{1}{n} + \frac{1}{m}))$
This means that we have the 95% confidence interval for $\mu_x - \mu_y$ as:
$\bar{X} - \bar{Y} \pm 1.96 * \sigma(\sqrt{\frac{1}{n} + \frac{1}{m}})$

15.2 Sec 11.1, 11.2

2 sample t-test with σ unknown

If we sample n times with replacement from $N(\mu_x, \sigma^2)$ and m times with replacement from $N(\mu_y, \sigma^2)$ then an unbiased estimator of σ is:

$$S_p^2 = \frac{(n-1)S_x^2 + (m-1)S_y^2}{(n-1) + (m-1)}$$

this is an unbiased estimator and:

$$\bar{X} - \bar{Y} = N(\mu_x - \mu_y, S_p^2(\frac{1}{m} + \frac{1}{n}))$$

Theorem 3 *This is the 2 sample t-test*
Suppose $X_1..X_n \sim N(\mu_x, \sigma_x^2)(i.i.d)$ *and* $Y_1, Y_2..Y_m \sim N(\mu_y, \sigma_y^2)(i.i.d)$
The test statistic given that $\sigma_x = \sigma_y$ *and that* S_p *is the standard error of both distributions combined:*

$$\frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2}$$

If we assume that these two have the same variance, but still unknown then we can use the combined variance estimate to plug in the bottom. However

if we assume that they are different the test statistic doesn't exactly abide by the t distribution but instead:

Theorem 4 *Suppose* $X_1..X_n \sim N(\mu_x, \sigma_x^2)(i.i.d)$ *and* $Y_1, Y_2..Y_m \sim N(\mu_y, \sigma_y^2)(i.i.d)$
The test statistic given that S_p *is the standard error is dependent on* $Var(\bar{X} - \bar{Y})$
$\frac{s_x^2}{n} + \frac{s_y^2}{m}$
Then we can approximate this as a t distribution with degrees of freedom equal to:

$$round(\frac{[(\frac{s_x^2}{n} + \frac{s_y^2}{m}]^2}{\frac{\frac{s_x^2}{n}^2}{n-1} + \frac{\frac{s_y^2}{m}^2}{m-1}})$$

Based on either, we have a corollary to the distribution showing the duality of confidence interval

Theorem 5 *A 100(1 - α)% CI for* $mu_x - \mu_y$ *is* $\bar{X} - \bar{Y} \pm t(\frac{\alpha}{2})S_{\bar{X}-\bar{Y}}$
Hypothesis testing for 2 *sample problem using* $t - test$ *is:*
$H_0 : \mu_x - \mu_y = 0$
$H_1 : \mu_x - \mu_y \neq 0$ *or*
$H_1 : \mu_x - \mu_y > 0$ *or*
$H_1 : \mu_x - \mu_y < 0$

If n, m are large the t test is approximately a $z - test$ and the t and z test are equivalent to the GLRT

16 Lecture 23

Last time: 2 sample t test
We compared the means of two independent normal populations:
same variance
$S^2_{\bar{X}-\bar{Y}} = (\frac{1}{n} + \frac{1}{m})(\frac{(n-1)S_X^2 + (m-1)S_y^2}{n+m-2})$
diff variance
$S^2_{\bar{X}-\bar{Y}} = \frac{S_x^2}{n} + \frac{S_y^2}{m}$

16.1 Power

We often see the highest power (lowest type 2 error) possible in an experiment

We know that power is a function of sample size; lets assume we have two independent populations with a Normal distribution and equal standard deviations and take two samples of the same size

$X_1..X_n \sim N(\mu_x, \sigma^2)$

$Y_1...Y_n \sim N(\mu_y, \sigma^2)$

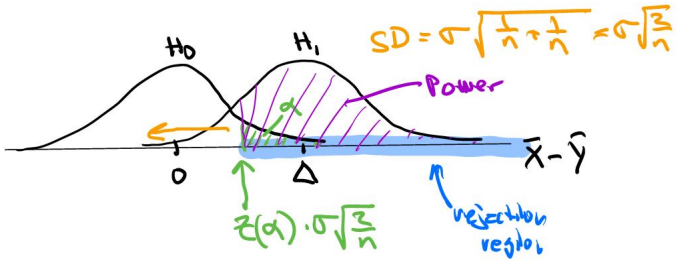
Hypothesis

$H_0 : \mu_x - \mu_y = 0$

$H_1 : \mu_x - \mu_y = \delta > 0$ (one sided)

If we know σ then we have:

$\bar{X} - \bar{Y} \sim N(\mu_x - \mu_y, \sigma^2(\frac{1}{n} + \frac{1}{m}))$



$$Power = P_1(d(x) = 1) \tag{2}$$

$$= P_1(\bar{X} - \bar{Y} > z(\alpha)\sigma(\sqrt{\frac{2}{n}})) \tag{3}$$

$$= P_1(\frac{\bar{X} - \bar{Y} - \delta}{\sqrt{\frac{2}{n}}} > \frac{z(\alpha)\sigma(\sqrt{\frac{2}{n}}) - \delta}{\sigma\sqrt{\frac{2}{n}}}) \tag{4}$$

$$= P_1(N(0,1) > \frac{z(\alpha)\sigma(\sqrt{\frac{2}{n}}) - \delta}{\sigma\sqrt{\frac{2}{n}}}) \tag{5}$$

$$\tag{6}$$

16.2 Paired t-test

We wish to estimate $\mu_x - \mu_y$ but in a paired study so our sample is no longer independent
let $X_1..X_n \sim^{iid} N(\mu_x, \sigma_x^2)$
let $Y_1...Y_n \sim^{iid} N(\mu_y, \sigma_y^2)$ $\sigma_{XY} = Cov(X_i, Y_i) \neq 0$
$D_i = X_i - Y_i$
$E(D_i) = E(X_i) - E(Y_i) = \mu_x - \mu_y$
$Var(D_i) = \sigma_x^2 + \sigma_y^2 - 2\sigma_{XY}$
$\bar{D} \approx N(\mu_x - \mu_y, \frac{1}{n}(\sigma_x^2 + \sigma_y^2 - 2\sigma_{XY})$

We take the test statistic:

$$t = \frac{\bar{X} - \mu_d}{S_{\bar{D}}} \sim t_{n-1}$$

$H_0 : \mu_d = 0$
$H_1 : \mu_d \neq 0$ and this has the rejection region:

$$|\bar{D}| > t_{n-1}(\frac{\alpha}{2})S_{\bar{D}}$$

16.3 Multinomial Distribution

Generalizes the binomial distribution
$p_1 \rightarrow$ outcome 1
$p_2 \rightarrow$ outcome 2
...
$p_m \rightarrow$ outcome m
$\sum_{i=1} p_i = 1$ We want to find the probability of getting x_1 of outcome 1, x_2 of outcome 2 ..
The answer of this is :

$$\binom{n!}{x_1!x_2!...x_m!} p_1^{x_1} p_2^{x_2} .. p_m^{x_m}$$

If we know what $p_1..p_m$ then the multinomial distribution is completely specified and we can calculate the probability of any outcome
We know that each x_i has the distribution of a $Bin(n, p_i)$ which is approximately normal $N(np_i, np_i(1-p_i))$ thus we can model our multionomial as

many normal distributions
We can see that this manifest as a Chi-Square distribution:

$$\sum_{i=1}^{m} \frac{(x_i - np_i)^2}{np_i} \sim \chi^2_{m-1}$$

$$\sum_{i=1}^{m} \frac{(x_i - np_i)^2}{np_i} = \sum_{i=1}^{m-1} \frac{(x_i - np_i)^2}{np_i(1-p_i)} = \sum_{i=1}^{m-1} \frac{(x_i - E(x_i))^2}{Var(x_i)} = \sum_{i=1}^{m-1} z_i^2$$

17 Lecture 24

17.1 Goodness of Fit χ^2 test

Our goal is to see whether our model for the distribution of a multinomial distribution fits our data
We draw n times with replacement from our box a get observed counts $x_1, x_2..x_m$ where $\sum_{i=1} x_i = n$. If the prob of tickets in box is $p_1(\theta)..p_m(\theta)$ we get expected counts $np_1(\theta), np_2(\theta)..np_m(\theta)$
Thus we do a goodness of fit test:

$$\text{Pearson Chi Square T.S} = \sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{m-1-k}$$

where k is the dimension of θ, O_i is the observed count and E_i is the expected count
A goodness of fit test explores how good your probability model fits your data

17.2 Hardy Weinberg (HW) Equilibrium Model

If gene frequencies are in equilibrium the genotypes AA, Aa, aa, occur in the population with probability: $(1 - theta)^2, 2\theta(1-\theta), \theta^2$ according to the HW model
Thus our null is that $P_1(\theta) = (1-\theta)^2, P_2(\theta) = 2\theta(1-\theta), P_3(\theta) = \theta^2$
We observe that $AA = 342$, $Aa = 500$,$aa = 187$

Step 1: Find $\hat{\theta}_{ML}$

$$lik(\theta) = \binom{n!}{x_1!x_2!..x_m!}(1-\theta)^{2x_1}(2\theta(1-\theta))^{x_2}\theta^{2x_3} \tag{7}$$

$$l(\theta) = log(n!) - \sum_{i=1}^{m} log(x_i!) + 2x_1 log((1-\theta)) + x_2 log(2\theta(1-\theta)) + 2x_3 log(\theta) \tag{8}$$

$$l'(\theta) = -\frac{2x_1}{1-\theta} + x_2\frac{2-4\theta}{2\theta(1-\theta)} + 2x_3\frac{1}{\theta} \tag{9}$$

$$0 = -\frac{2x_1}{1-\theta} + x_2\frac{2-4\theta}{2\theta(1-\theta)} + 2x_3\frac{1}{\theta} \tag{10}$$

$$0 = -4x_1(\theta) + x_2(2-4\theta) + 2x_3(2-2\theta) \tag{11}$$

$$0 = -4x_1\theta + 2x_2 - 4\theta x_2 + 4x_3 - 4x_3\theta \tag{12}$$

$$4\theta(x_1 + x_2 + x_3) = 2x_2 + 4x_3 \tag{13}$$

$$2\theta(n) = x_2 + 2x_3 \tag{14}$$

$$\theta = \frac{x_2 + 2x_3}{2n} \tag{15}$$

Step 2:
Thus, by this we get that:
$P_1(\hat{\theta}) = .320$
$P_2(\hat{\theta}) = .489$
$P_3(\hat{\theta}) = .180$
Thus our expected count should be: $x_1 = 3404, x_2 = 503, x_3 = 186$

Step 3:
Test:
H_0 : have Multinomial (1029, .32, .49, .18)
H_1 : have a MN of some other type with n = 1029 and 3 categories

$$\sum_{i=1}^{3}(\frac{(O_i - E_i)^2}{E_i}) = .0357$$

The probability of getting a number in a χ^2_{3-1-1} is about .85, thus it is above the α level.

18 Lecture 25

18.1 recap

The goodness of fit (G.O.F) χ^2 test assesses whether a categorical random variable follows a certain distribution (the null distribution) versus follows any other arbitrary distribution.

We saw that -2 log(Λ) where Λ is from the GLRT, is approximately the Pearson χ^2 test statistic

$$\sum_{i=1}^{\#cells} \frac{(O_i - E_i)^2}{E_i}$$

Since -2 log(Λ) $\sim \chi^2_{\text{dim sample space - dim null space}}$ it follows that

$$\sum_{i=1}^{\#cells} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{\text{dim sample space - dim null space}}$$

18.2 Assumptions of Goodness of Fit Test

- Have one categorical variable
- Have independent observations
- The outcomes are mutually exclusive
- We require large n and not many expected counts are below 5

We need all of these so that our multinomial is characterized by normal distributions and not by a Poisson or skewed normal

18.3 Test of Homogeneity

Example: A study was done comparing frequencies of a particular allele in a sample of diabetics and non diabetics
The following data was observed:

	Diab	non Diab
Bb or bb	12	4
BB	39	49

Our question is whether the non diabetic and the diabetic have differences in their frequency of alleles? **Theory**

J independent I cell multinomials
H_0 : all J multinomials are the same
H_1 : they are not all the same

19 Lecture 26

19.1 recap

The Test of Homeogenity is a way to assess if a J independent I cell multinomials all come from the same distribution?

19.2 Section 13.3

This is a continuation of the previous day:
Lets represent our data as such:

	Diab	No Diab	Total
Bb or bb	12	4	16
BB	39	49	88
Total	51	53	104

In a general case we can see that our observed values can be written as:

n_{11}	n_{12}	n_{13}	n_{1*}
n_{21}	n_{22}	n_{23}	n_{2*}
n_{*1}	n_{*2}	n_{*3}	n

The assumption we make is that the null distribution is $\begin{cases} \pi_1 \\ \pi_2 \end{cases}$

and thus our best estimates for these probabilities is:

:

$$\hat{\pi}_1 = \frac{n_{1*}}{n}$$
$$\hat{\pi}_2 = \frac{n_{2*}}{n}$$

Thus our expected counts are:

$n_{*1}\hat{\pi}_1$	$n_{*2}\hat{\pi}_1$	$n_{*3}\hat{\pi}_1$
$n_{*1}\hat{\pi}_2$	$n_{*2}\hat{\pi}_2$	$n_{*3}\hat{\pi}_2$

Thus the test statistic is the same, with:

$$\chi^2_{df} = \sum_{\text{all i, j}} \frac{(n_{ij} - \frac{n_{i*}*n_{*j}}{n})^2}{\frac{n_{i*}n_{*j}}{n}}$$

The next step is to figure out the degrees of freedom:

$$df = dim\Omega - dim w_0$$

w_0 is $I - 1$ because we can see this as a single multinomial with $I - 1$ degrees of freedom
Ω is equal to J multinomials each with $I - 1$ free params, thus we have $J(I - 1)$ - $I - 1$ which means we have $df = (J - 1)(I - 1)$

19.3 Test of Independence

Our goal is to see whether there two categorical variables are independent or not

We have J cell multinomial and
We have I cell multinomial
where both I and J are numbers (specifically positive integers)
$P(J = j) = \pi_{J=j} = \pi_j$
$P(I = i) = \pi_{I=i} = \pi_i$
$P(I = i, J = j) = \pi_{ij}$
Independence means that:
$\pi_{ij} = \pi_i \pi_j$

H_0 : Two independent R.V's size I, J
H_1 : Two dependent R.V's size I, J

Our observed and expected are exactly the same as the Test for Homogenity:

$n_{*1}\hat{\pi}_1$	$n_{*2}\hat{\pi}_1$	$n_{*3}\hat{\pi}_1$
$n_{*1}\hat{\pi}_2$	$n_{*2}\hat{\pi}_2$	$n_{*3}\hat{\pi}_2$

where $\hat{\pi}_{I=i} = \frac{n_{1*}}{n}$...
We get the same test statistic as the TOH

So, literally the difference between the Test of Independence and the Test of Homogenity is by design of experiment

In the test of independence, observational units are collected at random from the box and we observe 2 categorical variables for each of the elements. On the other hand, in the test of homogenity, the data is collected by random sampling from each subgroup of the subpopulation separately. Assumptions about the χ^2 test:

- Have one or two categorical variable
- Have independent observations or an SRS if it is relatively small to the sample size
- Outcomes are all mutually exclusive, thus we cannot have a member belonging to multiple values in the same categorical variable
- We require n to be large and no more than 20% of expected counts ¡ 5.

20 Lecture 27

20.1 recap

We have three different chi-square tests:

- GOF-test: have single sample from population , and answers whether the pop came from a particular multinomial distribution

- Test of Homogenity: multiple samples from subgroups of the population and answers whether the subgroups have the same multinomial distribution
- Test of Independece: we track two categorical traits from the same population, and answers the question if the categorical traits are dependent

These all run with the assumptions:

- Individuals in the population have a ticket that they can only mark one option on
- Tickets are drawn independently or a simple random sample that is a sample part of the total population of interest
- The numbers in the expected table aren't too small (less than 5)

20.2 Sec 11.2.3 Mann-Whitney, non parametric test

We use the Mann-Whitney test when we have 2 independent samples that aren't normal and we ask if the distributions are the same
$X = (x_1...x_m) \sim$ small group $= F($
$Y = (y_1...y_n) \sim$ large group $= G$

X and Y are independent samples
ex:
$x_1 = 1000, x_2 = 1380, x_3 = 1200, y_1 = 1400, y_2 = 1600, y_3 = 1180, y_4 = 1220$
Step 1: Sort $\{x_1...x_m, y_1..y_n\}$ from smallest to largest
$\{x_1, y_3, x_3, y_4, x_2, y_1, y_2\}$ We can then calculate T_X, T_Y, the rank sums of X, Y respectively. This is the Mann Whitney test statistic
$T_X = 9, T_Y = 19$
Null: $F = G$
Alternate: $F \neq G$

If the null is true that means that neither T_X or T_Y should be particularly large or small.

We need to produce some type of sampling distribution for T_X in order to evaluate the p- value of our calculated T_X. We can either:

- Permutate through all possible values of T_X given our sample. Thus there are $\binom{7}{3} T_X$'s and we can see where our values places in this; however, if we get really large m and n we are pretty much fucked.
- T_X and T_Y are approximately normal and use this fact to calculate.

$$T_Y \approx N(\frac{m(m+n+1)}{2}, \frac{mn(m+n+1)}{12})$$

This is proved using the Wilkoxin Test statistic:

$$U_Y = \sum_{i=1}^{n} \sum_{j=1}^{m} I(x_i < y_j) sim N(\frac{nm}{2}, \frac{mn(m+n+1)}{12})$$

The Wilkoxin test statistic has the following relationship with Mann Whitney test statistic:

$$U_Y = T_Y - \frac{m(m+1)}{2}$$

21 Lecture 28

21.1 recap

Mann Whitney Test - T_X and T_Y are the rank sums of X and Y. If the null is true, we expect that the rank sums will be close to each other. We can show that for large $m+n$ the distribution of the Mann Whitney test statistic is approximately normal:
Wilkoxin tried to estimate $\pi = P(X < Y)$ and in so doing showed the Mann Whitney test statistic T_X is approximately normal:

let

$$U_y = \sum_{i=1}^{m} \sum_{j=1}^{n} I(X_i < Y_j)$$

We know that this is approximately normal, by the CLT and the sum of independent variables.
To approximate π, Wilkoxin used $\hat{\pi} = \frac{1}{mn} U_y$

We know that $U_y = \sum_i \sum_j I(X_i < Y_j)$ is comparing every X to every Y. For any X_I we know that is is less than a given Y_i if it has a lower rank than the Y_i meaning we can rewrite this summation as:

$$U_y = \sum_i \sum_j I(X_{(i)} < Y_{(j)})$$

We know that T_y = the sum of the rank of Y. We can see that U_y is equal to:

$$\begin{gathered}(numberofX's) < Y_{(1)} = R_{(y_1)} - 1+ \\ (numberofX's) < Y_{(2)} = R_{(y_2)} - 2+ \\ .... \\ (numberofX's) < Y_{(m)} = R_{(y_m)} - m\end{gathered}$$

This means that $U_y = \sum_{i=1}^{m} R_{y_i} - \sum_{i=1}^{m} i$. This means that $U_y = T_y - \frac{m(m+1)}{2}$

We can also see that under the null $F = G$ that:

$$E(T_y) = \frac{m(m+n+1)}{2}$$

$$Var(T_y) = fracm(m+n+1)12$$

21.1.1 Proof

If we assume that $F = G$ then that means that T_y is equal to the sum of sampling m times without replacement from the population $\{1, 2...m+n\}$ Thus we know that:

$$E(T_y) = m\mu$$

$$Var(T_y) = m\sigma^2(\frac{N-m}{N-1})$$

Where μ and σ^2 are the population mean and variance of the average of a sample taken from population.

$$\mu = \frac{1}{N}\sum_{i=1}^{N} i = \frac{1}{N}\frac{N(N+1}{2} = \frac{N+1}{2}$$

$$\sigma^2 = E(X^2) - E[X]^2 = \frac{1}{N}\sum_{i=1}^{N} i^2 - (\frac{(N+1)}{2})^2$$

$$\sigma^2 = \frac{1}{N}\frac{N(N+1)(2N+1)}{6} - \frac{(N+1)^2}{4} = \frac{(N+1)(2N+1)}{6} - \frac{(N+1)^2}{4}$$

$$\sigma^2 = \frac{4N^2+6N+2-3N^2-6N-3}{12} = \frac{N^2-1}{12}$$

Thus we know that $(N = m + n)$:

$$E(T_y) = m * \mu = \frac{m(N+1)}{2} = \frac{m(m+n+1)}{2}$$

$$Var(T_y) = m\frac{N^2-1}{12} * \frac{N-m}{N-1} = \frac{(m+n+1)(mn)}{12}$$

22 Lecture 29

22.1 Recap

We can see that the Mann Whitney is 95% as accurate as a t-test for data assuming that the data is normal. Thus, the Mann Whitney will always preform close to the t-test and, is more flexible when we ae concerned with data in which we don't know the distribution exactly.

22.2 Signed Rank Test

This works for paired data. This shit is fat complicated but also sorta dope:

before	after	difference	abs(difference)	rank	signed rank
25	27	2	2	2	2
29	25	-4	4	3	-3
60	59	-1	1	1	-1
27	37	10	10	4	4

w_t (Wilkoxin Test Statistic) = sum of the positive signed ranks. H_0 We expect that the distribution of differences is symmetric around zero. For small n we can work out the different combinations and thus have the sampling distribution. For large n we use:

$$W_t \sim N(\frac{n(n+1)}{4}, \frac{n(n+1)(2n+1)}{24})$$

The Wilcoxin test is approximately 95% as powerful as the t-test but this is only when the assumptions of the t-test is meet.

22.3 Analysis of Variance

One way layout is an experimental design in which independent measurements are made under each of several treatments The model:

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

where ϵ_{ij} is a $N(0, \sigma^2)$ variable and $\sum_{i=1}^{n} a_i = 0$. These facts mean that: $Y_{ij} \sim N(\mu + \alpha_i, \sigma^2)$

$$H_0 : \alpha_1 = \alpha_2 .. = \alpha_n = 0$$

$$H_1 : \text{ Some alpha's will differ}$$

23 Lecture 30

23.1 recap

Analysis of variance is aimed at comparisons of the means of data.
The goal in a one way layout is to see whether the differences in the means of the measurements is significant or just due to chance.

23.2 Normal Theory; F-Test

Lets generalize and have I groups each with J observations, and we will call each of the elements in I as a "treatment". We will let:

$$Y_{ij} = \text{the j observation of the ith treatment}$$

and assume that our model is corrupted by independent random errors such that:

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

μ is the overall mean level, α is the differential effect of the ith treatment. The errors are assumed to be independent in the j^{th} observation of the i^{th} treatment. We assume that $\epsilon_{ij} = N(0, \sigma^2)$ and $\sum_{i=1}^{I} a_i = 0$ $E(Y_{ij}) = \mu + \alpha_i$ which means that if α_i is zero the value is the same for all the observations in that treatment. The difference between treatment i and j is is $\alpha_i - \alpha_j$

Analysis of variance is based on the following principle:

$$\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - \bar{Y}_{..})^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - \bar{Y}_{i.})^2 + \sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{Y}_{i.} - \bar{Y}_{..})^2$$

where

$$\bar{Y_{i.}} = \frac{1}{J}\sum_{j=1}^{J} Y_{ij}$$

and

$$\bar{Y_{..}} = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} Y_{ij}$$

The terms appearing in the first identity is called the sums of square and can be represented by:

$$SS_{TOT} = SS_W + SS_B$$

The sum of the squares equal the difference within groups and the difference between groups.

23.3 Independence of SSB and SSW

We know that $\bar{X}$ and $X_j - \bar{X}$ are independent R.V's. (If you don't remember the proof just accept it, its a whack a proof).

We know that $\bar{Y} = \frac{1}{I}\sum_{i=1}^{I} \bar{Y_{i.}}$ is a function of $\bar{Y_{1.}}, \bar{Y_{2.}}..\bar{Y_{I.}}$. This means that $\bar{Y_{i.}} - barY$ is also a function of $\bar{Y_{1.}}, \bar{Y_{2.}}..\bar{Y_{I.}}$. This also means that SSB is a function of $\bar{Y_{1.}}, \bar{Y_{2.}}..\bar{Y_{I.}}$. However, SSW is a function of $Y_{ij} - \bar{Y_{i.}} \forall i$ which means that by the fact that they either involve independent variables or by the theorem at the start of this section, these are all independent and thus SSB and SSW are independent.

23.4 Total Squares and Chi Square

$\frac{SST}{\sigma^2} \sim \chi^2_{IJ-1}, \frac{SSB}{\sigma^2} \sim \chi^2_{I(J-1)}$

If we use the fact that $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$ and $(n-1)S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2$ we can see that these are the same scenario.

NEEDS WORK Facts:

1. SST = SSW + SSB
2. SSW and SSB are independent

3. Chi square subtraction property: $X = X_1 + X_2, X \sim \chi^2_{n_1+n_2}, X_1 \sim \chi^2_{n_1} \to X_2 \sim \chi^2_{n_2}$ This means that SST has $IJ-1$ degrees of freedom, SSW has $I(J-1)$ degrees of freedom and SSB has $I-1$ degrees of freedom. (This makes sense since the total is looking at all the squares, within is comparing $J-1$ observatios in I groups and between is looking only between I treatments.

23.5 F-Distribution

Definition: Let U and V be independent χ^2 random variables with a and b degrees of freedom. The distribution

$$F_{a,b} = \frac{\frac{U}{a}}{\frac{V}{b}}$$

is called the F distribution.

The F-Test itself is:

$$F_{I-1,I(J-1)} = \frac{\frac{SSB/\sigma^2}{I-1}}{\frac{SSW/\sigma^2}{I(J-1)}} = \frac{SSB(I(J-1))}{SSW(I-1)}$$

24 Lecture 31

24.1 recap

Assume that we have a variety of treatments, each with a certain number of observations and we want to test if any of these had any statistically significant difference compared to the average.

We assume that each observations in the treatment behaves according to

$$Y_{ij} = \mu + a_j + \epsilon_{ij}$$

where the μ is the default we assume that all of them abide by and a_j is the impact of treatment j with an error that is normally distributed with zero mean and some standard deviation. Our goal is to see if all the a_j are zero or not. We can use a F test to see the ratio between the

SSB and SSW and use this to assess if the differences are statistically significant.

There are certain assumptions we make with the $F - Test$:

- Our data Y_{ij} is normal
- The variance of each treatment group is the same σ^2
- All observations are independent

If we dont meet this we should use a non parametric test

24.2 Review of T-test

Assume two independent normal populations $\mathcal{N}(\mu_x, \sigma^2)$ and $\mathcal{N}(\mu_y, \sigma^2)$ with the same variance. Take random samples $X_1, X_2..X_n$ are i.i.d $\sim \mathcal{N}(\mu_x, \sigma^2)$ and $Y_1...Y_n$ are i.i.d $\sim \mathcal{N}(\mu_y, \sigma^2)$

$$S_p^2 = \frac{(n-1)S_x^2 + (n-1)S_y^2}{2n-2}$$

$$\frac{(2n-2)S_p^2}{\sigma^2} = \frac{(n-1)S_X^2}{\sigma^2} + \frac{(n-1)S_y^2}{\sigma^2} = \chi_{n-1}^2 + \chi_{n-1}^2 = \chi_{2n-2}^2$$

due to the the standard errors being independent of each other.

Thus we see that:

$$t_{2n-2} = \frac{\bar{X} - \bar{Y} - 0}{SE(\bar{X} - \bar{Y})}$$

where the standard error is equal to $\sqrt{Var(\bar{X} - \bar{Y})} = \sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{n}} = \sigma * \sqrt{\frac{2}{n}} \approx S_p\sqrt{\frac{2}{n}}$

24.3 Bonferonni T-test

if the F-test rejects the null, it is likely that at least one pair of means is different. If we do $\binom{I}{2}$ pairwise t-test, at alpha level of α that means that the probabilit y that we expect at least one false positive is $\leq \alpha\binom{I}{2}$

by the addition rule (the probability each one is a false positive is α which means that the probability they aren't false positive is 1 - α which means the probability they are all not false positive is the sum of all of these substracted from 1). Using this we can set:

$\alpha = \frac{.05}{\binom{I}{2}}$

This means that as we get more tests it gets harder to reject a null value in each of the t-tests.

We can use Bonferroni t-test which uses the pooled sample variance:

$$S_p^2 = \frac{(J-1)S_1^2 + (J-1)S_2^2 ... + (J-1)S_I^2}{I(J-1)}$$

which has $I(J-1)$ df

Thus we have

$$t_{I(J-1)} = \frac{\bar{Y_{I*}} - \bar{Y_{J*}}}{S_p\sqrt{\frac{2}{J}}}$$

25 Lecture 32

25.1 Different size groups Anova

$$SST = \sum_{i=1}^{I}\sum_{j=1}^{J_i}(Y_{ij} - \bar{Y})^2, \frac{SST}{\sigma^2} \sim \chi^2_{\sum_{i=1}^{I} J_i - 1}$$

$$SSW = \sum_{i=1}^{I}\sum_{j=1}^{J_i}(Y_{ij} - \bar{Y}_{i*})^2, \frac{SSW}{\sigma^2} \sim \chi^2_{\sum_{i=1}^{I}(J_i - 1)}$$

$$SSB = \sum_{i=1}^{I}\sum_{j=1}^{J_i}(\bar{Y}_{i*} - \bar{Y})^2, \frac{SSB}{\sigma^2} \sim \chi^2_{I-1}$$

Bonferroni Tests

$$S_p^2 = \frac{SSW}{\sum_{i=1}^{I} J_i - 1}$$

$$t_{\sum_{i=1}^{I}(j_i-1)} = \frac{\bar{Y}_{i*} - \bar{Y}_{j*}}{S_p\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}}$$

and we compare this value to $\frac{\alpha}{\binom{I}{2}}$