551 Lecture Notes

Nick Sun

Lecture 1

- Experimental/Observational Units: smallest division of experimental elements under study. They usually receive different treatments.
- Variables: measured for each observational unit. Quantitative or Qualitative.
- Predictor/Independent/Covariate Variables are all used to classify experimental units and may be associated with the outcome variable of interest.
- Extending inference from sample to larger population involves sampling from those exact populations in some representative way.
- Inferences to populations can be drawn from random sampling studies, but not otherwise.
- Observational study (subjects choose to sit somewhere) vs. randomized experiment (subjects assigned to be in certain group)
- Statistical inferences of cause-and-effect relationship can be drawn from randomized experiments but not from observational studies
- Confounding variables; be sure to watch for the ecological fallacy! Relationships at the aggregated level may not exist at the individual level

Lecture 2

- Arithmetic mean, geometric mean, harmonic mean
- Population of interest, variable of interest, parameter
- Population distribution describes the range and relative likelihood of the set of possible values that Y can take on
- If Z has a Normal(0, 1) distribution, then X = $\sigma Z + \mu$ has a Normal(μ, σ^2)
- $Z = \frac{X-\mu}{\sigma}$ has a Normal(0, 1) distribution
- If X and Y are both normally distributed, the distribution of their sum is just adding their means and variances together
- The distribution of a statistics like $\bar{x}$ is referred to as the sampling distribution

Lecture 3

- $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$. This holds for any sample of size n from any population.
- We cannot obtain a sampling distribution of a statistics if we don't know the population distribution
- If we don't know pop distribution, we can do the following:
  - Work out certain properties (parameters) for example, the mean and variance of sampling distribution
  - Simulate
  - Approximate (using something like CLT)
- Remember that the mean is a linear operator E(X + Y) = E(X) + E(Y), don't have to be iid
- Var(Y) = $E\left[(X - E(X))^2\right] = E[X^2] + E[X]^2$

- Cov(X, Y) $= E\big[(X - E(X))(Y - E(Y))\big]$
  - If X and Y are independent, then Cov(X, Y) = 0, but the converse does not hold
  - Cov(X, X) = Var(X)
- For any two random variables X and Y, the variance of the sum is $Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y). A minus will just make the last term negative.
- Weak Law of Large Numbers
  - As sample size goes to infinity, the sample mean converges in probability to mean μ
  - $\bar{X} \rightarrow_p \mu$
- The ecdf of a function $\hat{F} = \frac{count\,obs\,less}{n}$
  - Can also be written $\hat{F}(x) = \frac{1}{n}\sum_{i=1} n 1_{x_i \leq x}$
  - Notice that its just a sample mean, so the weak law of large numbers applies
  - The ecdf converges to the true cumulative distribution function

Lecture 4

- Central Limit Theorem
  - If pop distribution of a variable X has population mean μ and a finite variance, then the sampling distribution of the sample mean goes closer to the Normal(μ, $\frac{\sigma^2}{n}$) as n increases
  - Equivalent statement: $\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \rightarrow_d N(0, 1)$
- $P(\bar{X} < 19.5) = P(\bar{X} - 20 < 19.5 - 20) = P\left(\frac{\bar{X}-20}{\sqrt{\frac{1}{4}}} < \frac{19.5-20}{\sqrt{\frac{1}{4}}}\right)$
  - Notice that the above LHS is now standardized and is distributed by N(0, 1)
- Good to know these R functions
  - dnorm(x, mu, sd) gives us the value of the density curve at x
  - pnorm(x, mu, sd) gives us the cumulative probability at x (basically our CDF in R)
  - qnorm(p, mu, sd) gives us the pth percentile (useful for finding critical values)

Lecture 5

- Example of sample size calculation such that if we know that $\mu = 20$ and $\sigma^2 = 4$, P(19.5 < $\bar{X}$ < 20.5) =.9
  - This can be found by knowing that we can restandardize all the values in the inequality
  - We know that 1.645 cuts off 5% in the tail for a normal distribution
  - So we just need to set both sides of the inequality to 1.645 and solve for n $\approx$ 44
- The process of using a sample to learn something about a population parameter is called inference
- What makes a good estimate?
  - Unbiased
  - Small mean squared error
  - Converges to true value as sample size increases (consistency)
- Null hypothesis: a specified value or range of values for the parameter of interest
- Alternative hypothesis: A different specified value of range of values for the parameter of interest
- We fail to reject the null hypothesis, we cannot prove that it is true

Lecture 6

- The rejection region of a hypothesis test is defined by the rejection distribution. It is the distribution to which the test statistic is compared and is considered under the null hypothesis being true
- Type I error is rejecting the null when the null is true
- Type II error is failing to reject the null when the null is false

- The significance level of a test is the probability of a type I error
- The power of a test at a specific value θ_A is the probability of rejecting the null knowing that the true value of the population parameter is θ_A.
  - This is equal to 1 - P(type II error)
- Rejection region is the values for which the null hypothesis will be rejected
- Using R, we don't have to standardize since we can just use qnorm() to get our critical values
- Usually though, we will just standardize the sample mean using the null hypothesis value of μ
- The *Z-test* is used to test hypotheses about a population mean when the population variance is known
  - Data setting is one sample, iid, with sample mean $\bar{X}$
  - the Null hypothesis is $H_0 : \mu = \mu_0$
  - Test statistic is $Z(\mu_0) = \frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}}$
  - $Z(\mu_0) \sim N(0, 1)$
- Exactness: under any setting for which the null hypothesis is true, is the actual rejection probability equal to the desired significance level α
  - Finite-sample exactness: For finite sample sizes n, is probability of rejecting the null = α where the null is true?
  - Asymptotic exactness: As the sample size goes to infinity, does probability of rejecting the null = α where the null is true?
- A test is finite sample exact if the reference distribution is the true distribution of the test statistic when the null hypothesis is true
- A test will be asymptotically exact if the reference distribution is asymptotic distribution of the test statistic when the null hypothesis is true
- Consistency: under any fixed setting for which the alternative hypothesis is true, does the rejection probability tend to 1 as the sample size goes to infinity
- The Z-test is finite sample exact if the data sampled is iid normal
- The Z-statistic is also asymptotically exact when the data sampled is iid (μ, σ^2), doesnt have to be normal

Lecture 7

- The power of the test if $\mu = \mu_A \neq \mu_0$ is given as

$$P\left(\frac{\bar{X}-\mu_A}{\sqrt{\frac{\sigma^2}{n}}} > z_{1-\alpha} + \frac{\mu_0-\mu_A}{\sqrt{\frac{\sigma^2}{n}}}\right)$$

Notice that we restandardized the LHS. It is distributed according to the N(0, 1)

- A p-value is the probability under the null hypothesis of observing a result at least as extreme as the statistic you observed
- A procedure for obtaining p-values is exact if the resulting value actually reflects the probability of obtaining results at least as extreme as the observed value under the null hypothesis
- Exact p-values under the null hypothesis should have a U(0, 1) distribution!
- Exactness of confidence intervals, p-values, and hypothesis tests are all dependent on the validity of the reference/null distribution
- Equivalent definition of a p-value: the lowest value of α for which the hypothesis test would be rejected
- The duality between hypothesis tests and confidence intervals:
  - A $(1-\alpha)100$ confidence interval is the set of all null hypotheses that would not be rejected at level α
  - A two-sided confidence interval corresponds to a two-sided alternative hypothesis test
- The formula for a confidence interval for a z-test is:

$$\bar{X} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma^2}{n}}$$

Lecture 8

- For a z-test, what do we do when we don't know σ^2? We can estimate it using the sample variance
- As the sample size n gets larger, the sample variance gets closer to the true population variance
- Replacing σ^2 for s^2, we get the t-test
- the null distribution of the t-statistic is the t-distribution, a family of distribution that are defined by a parameter called degrees of freedom
- The t-value $\frac{\bar{X}-\mu}{\sqrt{\frac{s^2}{n}}}$ has *exactly* a t_{n-1} distribution if the population distribution is exactly normal
- Otherwise, the t-statistic is asymptotically exact
- The R function for a t-distribution are: qt, pt, dt and they all take a df input
- Many of the formulas are similar to the z-distribution, but replace σ^2 with s^2
- Do not confused the sample variance, which estimates the population variance, with the $\text{Var}(\bar{X})$ which is often called standard error. We often estimate standard error by using s^2
- It is important to distinguish between *statistical significance* and *practical significance* when communiating results of an analysis
  - Statistical significance stems just from having a small enough p-value
  - Practical significance deals with effect size; does it have meaningful implications in real life?
  - It is possible to have practical significance without statistical significance (small n)
- When testing a binomial proportion, we have two options: use exact null distribution or use a normal approximation with a z-test

Lecture 9

Exact binomial test

```
+ one sided test where $H_{A}$ p > $p_{0}$
+ Given a binomial distribution, find the value c where P(X $\geq$ c) $\leq$ .05
+ This value c wont cut off .05 exactly unless you are very lucky
+ Specifying the lower tail alternative is very similar process
```

- For a two tailed exact binomial test, it's a bit different
  - what values do we reject for? Values that are far from expected value, or values that are least probable under the null hypothesis?
  - The former is the logic behind the z-test. The latter is the logic behind the binomial exact test
  - Look at the tails of the binomial distribution. Reject H_0 for X such that $p_0 \leq$ some value that is less than α
  - It may happen that we may be way below α but adding the next lower probability puts us over. We dont want our type I error probability to exceed α
    - Define our rejection region by $\sum_{k:p_0(k)\leq c} P_{H_0}(X=k) \leq \alpha$
    - Our p-value would be found by adding all the probabilities that are less likely or as likely as the observed x under the H_0 ### Z-test approximation for binomial response data
  - $X = \sum_{i=1}^{n} Y_i \ N(np, np(1-p))$ for large n
  - This gives us $\frac{X-np_0}{\sqrt{np_0(1-p_0)}} \ N(0,1)$
  - And a z-statistic that is $\frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}} \ N(0,1)$
  - Pretty much the same as the z-statistic that we already covered ### T-test for binomial test when we don't know the variance and estimate it using the s^2

- $\frac{\hat{p}-p_0}{\sqrt{\hat{p}(1-\hat{p})/(n-1)}}$
- The t-statistic converges to the z-statistic as n $\to \infty$
- The t-statistic here is basically the Wald statistic, just with an n-1 instead of an n
- The score test statistic ($z(p_0)$) performs slightly better than the Wald statistic ($z_W(p_0)$) in exactness and power
- For some reason, it is common to use a CI based on the Wald calculation of variance, even if the hypothesis test was done with the score test
- Use normal approximation when $np_0 > 5$ AND $n(1-p_0) > 5$

Lecture 10

- Sometimes folks will suggest using a randomized test with a binomial proportion so that P Type I error equals α exactly. It basically involves rejecting the borderline value only some probability γ of the time. You can find this using algebra, but we don't use randomized tests much ### Sign Test
  - Parameter of interest is the population median M
  - The sign test is basically just a binomial test where $H_0 : p_0 = .5$ where we are considering each observation a Bernoulli variable that is 1 if it is $\leq$ the median
  - Our test statistic is z = $\frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}}$
  - Keep in mind that our $p_0 = 0.5$ under the null
  - Recall that a $(1-\alpha)100$ confidence interval for a parameter θ is the set of all values for θ_0 for which a level α two sided test would not reject the null hypothesis
  - The way to find a CI for the population median is to find it iteratively using $\frac{\frac{X}{n}-0.5}{\sqrt{.5^2/n}} < z_{1-\frac{\alpha}{2}}$ where X is the number of observations less than M_0
  - Solving this, we get the interval $\frac{n \pm z_{1-\frac{\alpha}{2}}\sqrt{n}}{2}$. These indicate the ith observation which would fall into the confidence interval
  - The smallest observation is $\frac{n-z_{1-\frac{\alpha}{2}}\sqrt{n}}{2}$ and the largest is $\frac{n+z_{1-\frac{\alpha}{2}}\sqrt{n}}{2}+1$
  - we round to the nearest integer if the above are not integers
  - The CI bounds will always be values of the observed sample

Lecture 11

- The sign test for M is not finite sample exact because of the discrete nature of the data and we also use a normal approximation if we're using a z-test
  - The sign test will only be asymptotically normal
- The sign test is consistent. The test for binomial proportions is consistent as n approaches infinity, so the sign test follows similarly ### Wilcoxon Signed Rank test
  - Comes with a lot of caveats; be careful not to ignore assumptions for symmetric underlying distributions!
  - Definition as per lecture: The wilcoxon signed-rank test applies to the case of symmetric continuous distributions. Under this assumption, the mean equals the median. The null hypothesis is $H_0 : \mu = \mu_0$
  - Procedure
    * Calculate the distance of each observation from some proposed c_0
    * Rank the observations by the absolute value of their distance
    * Sum the ranks that correspond to observations larger than c_0
  - As before, we have a few options for a reference distribution: we an use an exact p-value by assuming each rank has the same chance of being above or below c_0 or we can use a normal approximation
  - If we use a normal approximation, we have sum S ~ $N(\frac{n(n+1)}{4}, \frac{n(n+1)(2n+1)}{24})$

- And our z-statistic will be $Z = \frac{S - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$
- The wilcox.test() function in R has options for using exact reference distribution and to use a continuity correction
- Only assumption for the Wilcoxon Signed Rank test is that the observations are independent of each other
- Does not tell us anything about the mean or median (actually tests the pseudomedian)

Lecture 12

- If we assume the underlying population is symmetric, the signed-rank test is a test of the population mean μ which is equal to median M which is also equal to the pseudomedian
- Consistent test of mean = median = pseudomedian under the symmetry assumption
- Not finite sample exact, but is asymptotically exact under the symmetry assumption
- For asymmetric distributions, not an exact test of the pseudomedian (but very close)
- For asymmetric distributions, test is still consistent for pseudomedian

One sample Chi-squared test for pop variance

```
+ Test statistic: $X(\sigma_{0}) = \frac{(n-1)s^{2}}{\sigma^{2}}$
+ The reference distribution for this test statistic is $\chi_{(n-1)}^{2}$
+ p-values are found using:
    + 1 - pchisq(X, n-1) for upper tail test
    + pchisq(X, n-1) for a lower tail test
    + 2*min(1 - pchisq(X, n-1), pchisq(X, n-1))
+ Confidence interval given as $\frac{(n-1(s^{2}))}{\chi^{2}_{(n-1), \alpha /2}}, \frac{(n-1(s^{2}))}{\c
+ If the underlying population distribution is not normal, this test is prettyyyy bad
```

- We can also do an *asymptotic t-test* for population variance
  - $t(\sigma_0^2) = \frac{\bar{Y} - \frac{n-1}{n}\sigma^2}{\sqrt{\frac{s_y^2}{n}}} \rightarrow_d N(0,1)$
  - We are essentially using a t-test to see if the population mean of Y_i is $\frac{n-1}{n}\sigma_0$

Lecture 13

Kolomogorov-Smirnov Test

- Say we want to test whether a population is distributed with a certain function F_X
- Our hypotheses are $H_0 : F = F_0$ and $H_A : F \neq F_0$
- Our test statistic is $D(F_0) = sup_x|\hat{F}(x) - F_0(x)|$ where $\hat{F}(x)$ is our empirical cumulative distribution function
- Our reference distribution is that under H_0, $\sqrt{n}D(F_0) \rightarrow_d K$ where K is the Kolmogorov distribution
- We are going to reject the null for high values of our test statistic $D(F_0)$
- $H_A : F > F_0$ implies that F is *stochastically smaller* than the null hypothesis F_0
- The one-sided test statistics are:
  - For $H_A : F > F_0$, $D(F_0) = sup_x\left(\hat{F}(x) - F_0(x)\right)$
  - For $H_A : F < F_0$, $D(F_0) = sup_x\left(F_0(x) - \hat{F}(x)\right)$
  - Be careful though! The interpretation of these tests is challenging. The two one-sided alternative hypotheses do not cover the full range of possibilities that could be going on here.
- The standard KS test should *not* be used if you are estimating parameters from the sample

- The KS test applies only to continuous distributions

Chi-squared Goodness of Fit test

- The discrete analogue of the Kolmogorov-Smirnov Test
- The test statistic is $X(p_0) = \sum_x \frac{n(\hat{p}(x)-p_0(x))^2}{p_0(x)}$
- The above is equivalent to how we usually see the test statistic written: $X(p_0) = \sum_{j=1}^{k} \frac{(O_j-E_j)^2}{E_j}$ where k is the discrete categories that the variable X_i can take on
- Suppose that we aren't specifying a distribution, but rather a family of distributions
  - We will have to estimate the parameters from the data, which we can do!
  - let's say that we are estimating d of these parameters
  - We use the null hypothesis with the estimated parameters and computer Pearsons χ^2 as usual
  - We compare the resulting statistic to a χ^2_{k-d-1} distribution where k is the number of categories and d is the number of estimated parameters
- The critical values for these tests can be found using qchisq function in R
  - For example, the critical value for a $\alpha = .05$ upper tail test with df = 5 is qchisq(.95, 5)

Lecture 14

- If we are given a discrete distribution with k-possible values and we want to test that P(X = x) = p_0, we can use Pearson's χ^2 test.
- For binary data, the χ^2 statistic is equal to the square of the z-statistic for testing a hypothesis for a binary proportion.
- Just like the case for binary data, the χ^2 distribution has an asymptotic χ^2 distribution.
  - This test is therefore asymptotically exact, but generally good when all expected values are over 5 (though this rule of thumb gets bent a lot)

Two-sample inference: The two sample z-test for population means

- Suppose we have two independent samples that may be from two different populations (different sample sizes too!)
- Always important to note the sampling context (are the sample sizes from each population fixed or are we doing an SRS from the combined population?)
  - When we analyze data that was gathered using an SRS, we can consider the proportion we get in our sample to estimate the true population proportion
- We can use a two sample z-statistic to test that the population means are different:
  - $H_0 : \mu_X = \mu_Y$ or alternatively but equivalently, $H_0 : \delta = 0$
  - $z(\delta_0) = \frac{\bar{X}-\bar{Y}}{\sqrt{\frac{\sigma_X^2}{m}+\frac{\sigma_Y^2}{n}}}$
  - Under the null, $z(\delta_0) \sim N(0, 1)$

Lecture 15

Slight detour into bootstrapping

- This a method to estimate a **nuisance parameter** which is something we aren't directly interested in but we need it to test the thing we are interested in
  - Classic example is something like population variance or sampling variance of a statistic
  - Basic idea is that the empirical distribution function converges to the true distribution function

  - Let's say we want to investigate medians for example. We have an initial sample. How do we get the distribution of the median in this population?
  - If we resample from our original sample (i.e. where we got our original empirical distribution) many times, we should get an idea of how this test statistic behaves
- Recall the important idea that as sample size increases, $\hat{F} \to F$
- Once we have our (probably thousands) bootstrap resamples with the associated test statistic of interest, we can calculate a boostrap confidence interval
- Suppose we have 1000 samples
  - The 95% CI is the 25th largest resample statistic and the 975th largest resampled statistic
  - Note that for some statistics, we might have a lot of duplicate values; this is alright

Lecture 16

Continuing with the two sample z-test for population means

- The rejection region for the z-test is similar to the standard z-test
- The confidence interval for δ_0 is $(\bar{X} - \bar{Y}) \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}$
- But what if we do not know the population variances? Common situation, glad you asked.
- Similar to the one sample case, we can just use the sample variance as estimates
  - **However**, we must consider two cases: when the population variances are equal and when they are unequal

T-test with an equal variance assumption

- Our best estimate of the population variance (and since the population variances are equal $\sigma_X^2 = \sigma_X^2 = \sigma^2$ is s_p
- Our pooled variance estimate is $s_p^2 = \frac{(m-1)s_X^2 + (n-1)s_Y^2}{m+n-2}$
  - Essentially we are finding a weighted average of the two sample variances; samples with more observations give better information about σ^2
- Our actual test statistic will look familiar:

$$t_E(\delta_0) = \frac{(\bar{X} - \bar{Y}) - \delta_0}{\sqrt{s_p^2\left(\frac{1}{m} + \frac{1}{n}\right)}}$$

- Under H_0 for normal populations, this t-statistic follows an exact distribution t_{m+n-2}
- As in the one-sample case, we still use the t-test even when we know that the data comes from non-normal populations - the t-test is pretty robust!
- And also for large sample sizes, deviations from normality don't make too much of a difference $m + n - 2 \to \infty$ then $t_{m+n-2}(p) \to z(p)$
- The confidence interval for this test, as you can imagine is just

$$(\bar{X} - \bar{Y}) \pm t_{m+n-2}(1 - \frac{\alpha}{2})\sqrt{s_p^2\left(\frac{1}{m} + \frac{1}{n}\right)}$$

- What happens when you use this test and the variances are actually not equal?
  - We see that the expected value of the estimated variance is **larger** than it should be when the **smaller** sample has the **smaller** variance
  - This makes sense; essentially we are not downweighting the variance estimate enough. We will reject less (less power)

- Conversely, we see that the expected value of the estimated variance is **smaller** than it should be when the **smaller** sample has the **larger** variance
  - Here, we are not downweighting the variance estimate too much! We will reject more (more power)

T-test with an unequal variance assumption

- If we make the unequal variance assumption, we find that the test statistic is actually not too bad looking, in fact its a familiar friend

$$t_U(\delta_0) = \frac{(\bar{X} - barY) - \delta_0}{\sqrt{\frac{s_X^2}{m} + \frac{s_Y^2}{n}}}$$

- However, we do not get an exact distribution for this test even if the populations are Normal because the denominator is not the square root of a chi-squared r.v. divided by its df
- Therefore, we have the use either the asymptotic Normal reference distribution (not great) or the Welch-Satterthwaite approximation
  - Essentially, we are saying that $t_U(\delta_0) \sim t_\nu$ where ν is this ugly damn thing:

$$\nu = \frac{W^2}{M + N}$$

where $W = \left(\frac{s_X^2}{m} + \frac{s_Y^2}{n}\right)^2$, $M = \frac{s_X^4}{m^2(m-1)}$, $N = \frac{s_Y^4}{n^2(n-1)}$

- The CI is almost the same as above

$$(\bar{X} - \bar{Y}) \pm t_\nu(1 - \frac{\alpha}{2})\sqrt{\left(\frac{s_X^2}{m} + \frac{s_Y^2}{n}\right)}$$

- Equal-variance t-test vs Unequal-variance (Welch) t-test: when sample sizes are equal, both test statistics are the same but the (degrees of freedom for the reference distributions still differ)
- When variances are equal, the equal variance t-test has slightly better power and slightly better exactness
- For unequal sample szies with unequal population variances, equal-variance t-test does not have the correct calibration!

Lecture 17

Paired data z-test

- Here we are supposing that the data is coming in pairs (before and after perhaps, or maybe siblings? Lots of possible scenarios)
- The two samples are by necessity the same size here
- Suppose we are interested if the difference in population averages is equal to 0
- We can acutally do a few different things here; we can look at the difference between the sample averages or look at the pairwise differences. They are equivalent.
- Instinct tells us that we should use a two-sample z-test here, but note that $\bar{X}$ and $\bar{Y}$ are not independent here! There is a covariance factor that we need to account for
  - $Var(\bar{X}, \bar{Y}) = \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{n} - 2\frac{\sigma_{XY}^2}{n}$
- Our z-statistic is therefore

$$z(\delta_0) = \frac{(\bar{X} - \bar{Y}) - \delta_0}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{n} - 2\frac{\sigma_{XY}^2}{n}}}$$

- The CI as you can imagine is

$$(\bar{X} - \bar{Y}) \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{n} - 2\frac{\sigma_{XY}^2}{n}}$$

- But also as you know we usually do not know the population variances. Therefore, we have to estimate them from the data.
- Our test statistic in this case is

$$t_(\delta_0) = \frac{(\bar{X} - \bar{Y}) - \delta_0}{\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{n} - 2\frac{s_{XY}^2}{n}}}$$

- If the differences D_i are notmal, then the t-statistic has an exact t-distribution with n-1 degrees of freedom
- Important to note: the Normality of X and Y does not imply the normality of D unless (X, Y) are jointly multivariate normal
- To recap:
  - Take the differences $D_i = X_i - Y_i$
  - Perform a one-sample hypothesis test for the population mean difference $\mu_d = \mu_X - \mu_Y$

Lecture 18

For the next 2 lectures, we are looking at a 2x2 contingency table. We are going to use the following convention:

	0	1	
X_i	a	b	m = a+b
Y_i	c	d	n = c+d
	a+c	b+d	N = a+b+c+d

- If we are given a 2x2 contingency table, more often than not we are interested in seeing if the probabilities $p_x = p_y$
  - We may also be interested in differences in proportion (subtracting $p_x - p_y$), relative risk ($\frac{p_x}{p_y}$), or the odds ratio (odds of x over odds of y)
  - Note that odds is calculated $\frac{p_X}{1-p_X}$
  - In the above table, odds ratio can be easily calculated by the following formula $\frac{ad}{bc}$

Two sample z-test of binomial proportion

- The z-statistic is calculated easily by

$$z = \frac{\hat{p}_x - \hat{p}_y}{\sqrt{\hat{p}_c(1 - \hat{p}_c)(\frac{1}{m} + \frac{1}{n})}}$$

Where $p_c = \frac{b+d}{N}$ and $Z \sim N(0, 1)$

- And the CI is found by

$$\hat{p_x} - \hat{p_y} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p_x}(1-\hat{p_x})}{m} \frac{\hat{p_y}(1-\hat{p_y})}{n}}$$

- Notice that this a wald interval with a score z-statistic!

Lecture 19

Chi-squared test for the homogeneity of proportions

- The chi-squared statistic is calculated as expected:

$$\chi = \sum \frac{(Obs - Exp)^2}{Exp}$$

- The expected values of the tables is found using the following calculations:

m($\frac{a+c}{N}$)	m($\frac{b+d}{N}$)	m
n($\frac{a+c}{N}$)	n($\frac{b+d}{N}$)	n
a+c	b+d	N

- Reject H_0 if $\chi^2 > \chi_1^2(1-\alpha)$

Fisher Exact Test

- The "test statistic" here is us the probability of us getting our observed table conditioned on the margins of the table
- The p-value is the sum of the probabilities of all the tables more extreme than the observed table
- The exact calculation of the probability of the observed table models after the hypergeometric distribution:

$$\frac{\binom{a+c}{a}\binom{b+d}{b}}{\binom{N}{a+c}}$$

- The definition of more extreme depends on the alternative hypothesis:
  - If $H_A : p_x > p_y$, more extreme means bigger values of $Obs_{1,2}$
  - If $H_A : p_x < p_y$, more extreme means smaller values of $Obs_{1,2}$
  - If $H_A : p_x \neq p_y$, more extreme means less likely table than our observed
- It's obviously tedious to calculate all of the tables so we usually just let the computer do it lmao

Quick aside on sampling

- Multinomial sampling is when we get N experimental units, classify each according to a grouping variable G and response variable X
- Two-sample Binomial sampling is when we obtain fixed sizes of m and n from each group
- For rare events, it can be challenging to obtain data using multinomial or two-sample binomial sampling
  - Example: getting people who are struck by lightning might be hard because, well, not many people get struck by lightning
  - We might decide in that case to just sample from people that we know got struck by lightning as one of our groups
  - Note that we can no longer estimate P(Lightning | Golfer), but we can still estimate P(Golfer | Lightning) and the odds ratio
  - The odds ratio fact is useful because we can flip the conditionals with odd ratios (useful property) so we can still say something about P(Lightning | Golfer)

Lecture 20

Log-odds Ratio Test

- Our test statistic here is the sample odds ratio $\hat{\omega} = \frac{ad}{bc}$
- The log of this estimate is asymptotically normal $log(\hat{\omega}) \sim N(log\omega, \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d})$
- To test $H_0 : \omega = 1$ we can use the following test statistic:

$$Z = \frac{log(\hat{\omega})}{\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}}$$

- $Z \sim N(0, 1)$
- As you can imagine, you can also create a CI for log(w) using the following:

$$log(\hat{\omega}) \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

- You can find a CI for ω just by exponentiating everything above
- Let's say that our sample odds ratio came out to be 1.667. This tells us that odds of getting struck by lightning are 1.667 times higher if you golf.
- You can perform the log-odds ratio test no matter how the data was sampled
  - Test performance will be better for large sample sizes
- You can also use Pearson's χ^2 test and Fisher's exact test no matter how the data was sampled, since all the tests assess whether there is an association between the variables.
  - Only the estimates are affected by the sampling scheme

Lecture 21

Paired Binary Data: McNemar's Test

- These are typically before and after type studies where the response is binary (ex: political opinion before and after a debate) or matched case-control sampling
- In these scenarios, we generally want to test the null hypothesis that $H_0 : p_{before} = p_{after}$
- It would not be appropriate to use two-sample binomial z-test, Pearsons, Fishers exact test, or log-odds ratio test here because they ignore the pairing information
- In this case, we should treat each pair as a single entity and our contingency table will be the counts of pairs
- In this case we can use McNemar's Test which is actually equivalent to the paired t-test in the sense that the test statistics are monotone transforms of one another
- In McNemar's Test we condition on the number of discordant pairs, i.e. pairs where the values don't match i.e.i.e. b + c
- Under the null hypothesis, half of b+c should be in $O_{1,2}$ and the other half should be in $O_{2,1}$
- Our test statistic is therefore the following:

$$z = \frac{b - c}{\sqrt{b - c}}$$

- Under the null hypothesis, $z \sim N(0, 1)$
- Equivalently, you can square this statistic and then compare it to χ^2_1 but its the same thing don't kill yourself
- Note that in this setting, the question we are still asking is "Is being struck by lightning associated with golfing?" in the case of us looking at siblings who golf

- The question of whether or not the status of the lightning struck sibling is independent of the other sibling can be answered with Pearson's, Fisher's, etc.
- For paired t-test vs McNemar's Test with large sample sizes, essentially, they are asymptotically equivalent! Their statistics tend to the same value

Lecture 23

Wilcoxon Rank-Sum Test or the Mann-Whitney U test

- **Not** actually a test of population medians as you were led to believe in undergrad. This is only true under strong assumptions with the population distributions
- We calculate the U statistic by:
  - Combining the two samples
  - Ranking the observations in the combined sample from smallest to largest
  - Add up the ranks corresponding to the observations in the smaller of the two groups
- There are a few ways to get p-values:
  - One is to use a **permutation** approach i.e. if there were **no** difference between the two groups populations, then each rank between 1 and $(n_x + n_y)$ will have the same chance of being assigned to the smaller group
  - This is computationally intensive though because to get an exact p-value you have to calculate $\binom{n_x+n_y}{n_y}$ total Rank statistics. This can get out of hand quickly!
  - Once we have a reference distribution for the U statistic, we can see where our observed U statistic falls in that distribution and calculate a p-value
  - The other practical way is to use a **normal approximation**
  - $R \sim N(E(R), Var(R))$ where $E[R] = \frac{n_x(n_x+n_y+1)}{2}$ and $Var[R] = \frac{n_x n_y(n_x+n_y+1)}{12}$
  - Our test statistic in this case would be

$$Z = \frac{R - E[R]}{\sqrt{Var[R]}} \sim N(0, 1)$$

- Some issues that arise in the Wilcoxon Rank-Sum test:
  - Ties in observed values
    * Assign ranks to observations as usual, then average the ranks assigned to tied values
    * Permutation approach to calculate p-values still works, but tabled values will not be correct since they assume no ties
    * If number of tires is large relative to the sample size, the normal approximation will not be very good
  - Normal approximation in small sample sizes
    * Basically adding .5 to the observed value of R if you are computing a lower probability and subtract .5 from the observed value of R if you are computing an upper probability
    * Slightly improves approximation for small sample sizes
  - Proper interpretation of the test results
    * Some sources describe the Wilcoxon Rank-Sum test as a test for an additive effect (essentially a shift between distributions; shapes and scales do not change at all)
    * If you are willing to assume that the only difference between populations is a shift, then you can use Wilcoxon Rank-Sum to test whether the shift is 0
    * If you are **not** willing to assume that the only difference is a shift, the interpretation of the Wilcoxon Rank-Sum test is
      - The test is that $H_0 : P(X > Y) = .5$ where X is a randomly chose value from population 1 and Y is a randomly chosen value from population 2
    * If you are assuming an additive effect, then the Wilcoxon Rank-Sum test is a test in difference in medians (but also means, percentiles, maxima, minima, etc.)

        * If you are not assuming an additive effect, the Wilcoxon Rank-Sum test does not say anything about medians
        * The Wilcoxon Rank-Sum test is an exact test of $H_0 : F_X = F_Y$ but is not exact in testing medians, means, or $P(X > Y) = .5$ unless we are assuming location shift
        * The Wilcoxon Rank-Sum test is not a consistent test for medians, means, or equality of distribution unless we are assuming a location shift, but it is a consistent test of $H_0 : P(X > Y) = .5$

Lecture 24

Two sample inference for population medians (Mood's Test)

- Here we are testing $H_0 : m_X = m_Y = m$.
- $\hat{m}$ is an unbiased and consistent estimator for the population median ($\hat{m} \sim N(m, \frac{1}{4nf(m)^2})$)
- Mood's Test procedure:
    - Find combined sample median $\hat{m}$
    - Test that the true population proportion of Xs greater than $\hat{m}$ is equal to the true population proportion of Ys greater than $\hat{m}$.
    - Once we have a new 2x2 table, we can use a two sample binomial proportion z-test or Pearson's chi-squared test or Fisher's exact test

Permutation tests

- General procedure:
    - Select a test statistic W that measures the kind of difference you are interested in measuring between two populations
    - Permute the group labels among observations and recalcualte test statistic
    - Repeat many times to obtain a permutation distribution for the test statistic
    - Calculate p-values against this permuation distribution
- Depending on the test statistic, the performance of the permutation test can vary
    - In most settings, the test detects **more** than the comparison indicated by the test statistic
    - Ex: the test will not reject at the correct rate if the population medians are equal but the population distributions differ
    - This stems from the assumption that the observations from the two populations are exchangeable (i.e. same population distribution, not just the individual statistic you are interested in)

Lecture 25

Two-sample inference for population variances

- We are interested in testing the equality of variances between two populations
    - Our underlying assumption here is that the populations are Normal
    - Recall that sample variances are unbiased and consistent estimators for the population variance and that $\frac{(n-1)s_X^2}{\sigma_X^2} \sim \chi^2_{(n-1)}$
- Our null hypothesis is typically $H_0 : \sigma_X^2 = \sigma_Y^2$ or more generally $H_0 : \sigma_X^2 \ \sigma_Y^2 = r$
- Our test statistic is the F statistic:

$$F(r) = \frac{s_x^2/\sigma_x^2}{s_y^2/\sigma_y^2} = \frac{s_x^2}{s_y^2}\left(\frac{1}{r}\right)$$

- Under the null hypothesis, $F(r) \sim F_{(m-1,n-1)}$
- This test does not perform well if the underlying population distribution is not Normal
  - The test will not reject with the correct probability when the null hypothesis is true (not exact or asymptotically exact)
  - The test is consistent, but the lack of calibration means it is hard to interpret the results
  - It is important to use this only when Normality of both populations is known

Levene's Test

- We are still testing equality of popuation variances here $H_0 : \sigma_X^2 = \sigma_Y^2$
- Test procedure
  - We can construct a new variable measuring the absolute difference of each observation from its sample median
  - $U_i = |X_i - med(X)|$ and $V_i = |Y_i - med(Y)|$
  - Perform a two sample t-test to test the hypothesis that the population mean of the U_i is the same as the population mean of the V_i
  - We can also used the squared differences instead of absolute value, or take the differences from the sample mean instead of the median
    * Option 1 is illustrated above: absolute difference with sample median
    * Option 2 is squared difference with sample median
    * Option 3 is the absolute difference with the sample mean
    * Option 4 is the squared difference with the sample mean
  - Welch's t-test is more robust then the t-test with equal variances
- The interpretation of Levene's Test depends on the option used
  - If you use option 4, you can interpret this as testing a difference in population variances
  - For the other options, this test is not using familiar quantities
- Assumptions:
  - Independence between samples
  - Reasonably large sample sizes so we can get lots of Us and Vs
  - If we use the equal-variance t-test at the end, we are automatically assuming that U and V have equal population variances
- Used to answer direct questions about variance and spread
- The R package 'car' has a leveneTest() function but that function only uses absolute value options

Two sample Kolmogorov-Smirnov Test

- We might want to test equality of the entire distribution function as opposed to just specific quanitites
- Our $H_0 : F_X = F_Y$ is similar to the one sample case
- Our test statistic D is also pretty similar to the one sample case

$$D = sup_x|\hat{F}_x(x) - \hat{F}_y(y)|$$

and our reference distribution is the Kolomogorov distribution

$$\sqrt{\frac{mn}{m+n}}D \to_d K$$

where we reject for large values of $\sqrt{\frac{mn}{m+n}}D$

- As before, the KS test applies only to continuous distributions
- If we want to test discrete distributions, we can use Pearson's Chi-squared test for r x c contingency tables

Delta method

- Used to approximate the sampling distribution of a function of a statistic whose asymptotic distribution is known
- For example, using the central limit theorem we know that $\sqrt{n}(\bar{X} - \mu) \rightarrow_d N(0, \sigma^2)$
- Suppose we are interested in the distribution of $\bar{X}^2$, so $g(x) = x^2$
- If we have a statistic T s.t. $\sqrt{n}(T - \theta) \rightarrow_d N(0, \tau^2)$ then for any continuous function g s.t. g' exists, we have

$$\sqrt{n}(g(T) - g(\theta)) \rightarrow_d N(0, \tau^2[g`(\theta)]^2)$$

Or in other words

$$g(T) \sim N(g(\theta), \frac{\tau^2[g(\theta)]^2}{n})$$

- This comes from the Taylor expansion
- The delta method provides estimates of the mean and variance of the function of a statistic:
  - $E[g(T)] \approx g(E[T])$
  - $Var[g(T)] = Var[T][g'(\theta)]^2$
  - These approximations can get pretty rough and in general unless g is a linear function the expectation of a function does not equal the function of the expectation
- Some sources recommend transforming data to improve the approximation of normality (reduce asymmetry) and make the Normal-based methods perform more exactly
  - Testing hypotheses regarding population means on orignal data can answer a **different** question than testing on transformed data
  - **However**, transforming inference back to the original scale is very challenging to interprety unless strong assumptions are made

Lecture 26 (Last lecture of the quarter!)

Mantel-Haenszel Test

- The setting here is k 2x2 tables under different conditions
- Our null hypothesis is that $H_0 : p_{xj} = p_{yj}$ for all j from 1 to k
  - Notation is $p_{xj} = P(X = 1 intable j)$
- Often expressed in terms of the odds ratio:
  - $\omega_j = \frac{\frac{p_{xj}}{1-p_{xj}}}{\frac{p_{yj}}{1-p_{yj}}}$
  - $H_0 : \omega_j = 1$ for all j from 1 to k
- Example scenario: is political preference associated with level of education
  - We could collect data from each state and each state would be a 2x2 table
  - In other words, we are asking: is the probability of being a Democray the same for people with and without a college degree in each state?

- We cannot combine the tables together into one; run the risk of Simpson's paradox

- Mantel-Haenszel test procedure:
  - $H_0 : \omega_j = 1$ for all j from 1 to k
  - $E[n_{x1j}] = \frac{n_{x.j}n_{.1j}}{n_{..j}}$
  - $Var[n_{x1j}] = \frac{n_{x.j}n_{y.j}n_{.1j}n_{.0j}}{n^2_{..j}(n_{..j}-1)}$
  - and our test statistic C is

$$C = \frac{\sum_j (n_{x1j} - \mu_{x1j})^2}{\sum_j \sigma^2_{x1j}}$$

and under H_0, $C \sim \chi^2_1$ and we reject H_0 for large values of C (p-value is 1 - pchisq(C, 1))

- The Mantel-Haenszel test assumes that odds-ratios are the same in all k tables
  - If this assumption is not met, it is difficult to interpret a p-value
  - The test may **fail** to reject the null if the odds ratio are different from 1 but in opposite direction