Lab #11

We will explore observational studies in today's lab and review how to make inferences on contingency tables. We will only use 2x2 tables for today's lab. Then we will review for the quiz.

1. χ^2 and Fisher's Exact Test Review

The χ^2 test and Fisher's Exact Test both do the same thing: they test whether or not two groups are dependent. As in, does being in one group for variable *A* affect whether you will be in a certain group for variable *B?* The χ^2 test is approximate while the Fisher's Exact test is exact.

<div align="center">

Variable B

		Y	N	
Variable A	Y	*a*	*b*	*a+b*
	N	*c*	*d*	*c+d*
		a+c	*b+d*	*N = a+b+c+d*

</div>

The χ^2 test works when all *expected* cell counts are not near zero, since it is an approximate test. Usually the Fisher's Exact Test results are about the same as the χ^2 test for 2x2 contingency tables; however, you must make sure expected cell counts under the null hypothesis of independence is not near zero (about greater than 5).

2. Odds Ratio and Relative Risk

Refer to the table above for this section.

The **odds of an event is** the ratio of the number of times the event occurs to the number of times the event fails to occur.

The odds for an event for the first row and second row are *a/b* and *c/d*, respectively. Since the odds ratio is literally the ratio of two odds we have:

$$\widehat{OR} = \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{ad}{bc}$$

What does an odds ratio of 1 mean?

The **risk** of an event is the probability of an event occurring. Therefore, the relative risk is the ratio of the risks for two groups.

$$\widehat{RR} = \frac{a/(a+b)}{c/(c+d)}$$

Estimated x% confidence interval for the **true log(Odds Ratio)** of the population of interest:

$$\log(\widehat{OR}) \pm Z_{x\%}\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

NOTE: This is the confidence interval for the log(Odds Ratio). So we will have to exponentiate it to get a confidence interval for the actual Odds Ratio.

X% Confidence interval: exp(lower bound for log OR), exp(upper bound for log OR).

3. Example

EFM Exposure	Caesarian Delivery		Total
	Yes	**No**	
Yes	358	229	587
No	2492	2745	5237
Total	2850	2974	5824

a) What is the estimated relative risk?

RR = (358/587)/(2492/5237) = 1.282

b) What is the estimated odds ratio?
(358*2745)/(229*2492) = 1.72

c) Find the 95% confidence interval for the true odds ratio. Is one inside of your interval? What do you conclude?

log(1.72) +/- 1.96*sqrt(1/358+1/229+1/2492+1/2745) = (0.37, 0.72).
exp(0.37), exp(0.72) = (1.45, 2.05)

Quiz 3 Review

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

BINOMIAL DISTRIBUTION

Uses: This is for **EXACT** probability. This finds the probability, given n, k, and p, of observing EXACTLY k events.

Useful note: $P(X = 0) + P(X = 1) + \ldots + P(X = n) = 1$. Let 1 represent this space. For at least one event:

$1 - P(X = 0) = P(X = 1) + P(X = 2) + \ldots + P(X = n)$

$P(X \leq k) = P(X = 0) + P(X = 1) \ldots P(X = k)$

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

APPROXIMATE NORMAL FOR BINOMIAL DISTRIBUTION

Uses: for hypothesis testing whether the true proportion is equal to p_0. Under the null hypothesis, this statistic follows an approximate normal distribution.

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

NORMAL DISTRIBUTION

Uses: For finding probabilities/proportions in normally distributed data

---

$$T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

STUDENT'S T DISTRIBUTION

Uses: testing whether the true mean of a population is equal to μ_0.

$$T = \frac{\overline{x_d} - 0}{s_d/\sqrt{n}}$$

Paired data: Paired T-test

Uses: Testing if the true mean difference between paired data is equal to zero. Uses the magnitude of the difference and is generally more powerful than the binomial test.

1) Record how many pairs had one treatment that did better than the other. Count as 'successes'.
2) Find the probability of observing data as extreme or more extreme, assuming the null hypothesis to be p = 0.5.

Paired data: Binomial Test

Uses: Assuming 50-50 chance of a treatment being better than another, treat as binomial distribution with probability = 0.5. Does not take into effect the magnitude of the difference of treatments.

POWER: probability of rejecting the null hypothesis when the null hypothesis is in reality false.
Affected by sample size, effect size, and variability.

CONFIDENCE INTERVALS

$$\hat{p} \pm Z_{X\%}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$
$$\bar{x} \pm T_{x\%,df} \; s/\sqrt{n}$$

These are only a select few of the topics covered for the material for the quiz on Thursday. Please review all lecture notes and homework solutions, especially the procedure slides in the lecture notes.

PRACTICE PROBLEMS

1. A company owns 400 laptops. Each laptop has an 8% probability of not working. You randomly select 20 laptops for your salespeople.

   (a) What is the probability that 5 will be broken?

   $P(X = 5) = \binom{20}{5}0.08^5(0.92)^{15} = 0.015$

   (b) What is the probability that they will all work?

$P(X = 0) = \binom{20}{0}0.08^0(0.92)^{20} = 0.19$

   (c) What is the probability of at most two not working?

$P(X <= 2) = \binom{20}{0}0.08^0(0.92)^{20} + \binom{20}{1}0.08^1(0.92)^{19} + \binom{20}{2}0.08^2(0.92)^{18} = 0.79$

2. We are interested in testing whether a certain at-risk population for diabetes has a daily sugar intake that is equal to the general population, which is equal to 77 grams/day. A sample of size 37 was taken from this at-risk population, and we obtained a sample mean of 80 and sample standard deviation of 11 grams.

   (a) Perform a hypothesis test to test whether this population has a significantly different mean sugar intake from 77 grams.

   $T = \dfrac{80-77}{(11/\sqrt{37})} = 1.66$, df = 36. Use table at df = 35. Pvalue between 0.15 and 0.1. Probably about 0.12. Exact p-value is:

```
> (1-pt(1.66, df =36))*2
[1] 0.1056019
```

   We fail to reject the null hypothesis and cannot conclude this at-risk population mean sugar intake is different from the general population mean sugar intake (p = 0.11)

3. The distribution of LDL cholesterol levels in a certain population is approximately normal with mean 90 mg/dl and standard deviation 8 mg/dl.

   (a) What is the probability an individual will have a LDL cholesterol level above 95 mg/dl?

   $Z = (95-90)/8 = 0.625$. $P(Z > 0.625) = 0.27$

(b) Suppose we have a sample of 10 people from this population. What is the probability of exactly 3 of them being above 95 mg/dl?

$$P(X = 3) = \binom{10}{3}0.27^3(0.73)^7 = 0.261$$

(c) Take the sample of size from part b. What is the probability that their sample mean will be above 95 mg/dl?

$$Z = \frac{95-90}{8/\sqrt{10}} = 1.98\ldots P(Z>1.98) = 0.024$$

(d) Suppose we take 5samples of size 10 from the population. What is the probability that at least one of the sample means will be greater than 95 mg/dl?

$$P(X >= 1) = 1 - P(X = 0) = 1 - \binom{5}{0}0.024^0(0.976)^5 = 0.11$$

4. A study indicates that 4% of American teenagers have tattoos. You randomly sample 30 teenagers. What is the probability that exactly 3 will have a tattoo?

$$P(X = 3) = \binom{30}{3}0.04^3(0.96)^{27} = 0.09$$

5. An XYZ cell phone is made from 55 components. Each component has a .002 probability of being defective. What is the probability that at least one XYZ cell phone will not work perfectly?

$$P(X >= 1) = 1 - P(X = 0) = 1 - \binom{55}{0}0.002^0(0.998)^{55} = 0.104$$

6. In the following scenarios, identify what will happen to the power of a hypothesis test:

   (a) We increase the sample size.
   Power increases

   (b) The standard deviation of the sample is larger than what we expected.
   Power decreases

   (c) Our effect size moves from 5 units to 10 units.
   Power increases