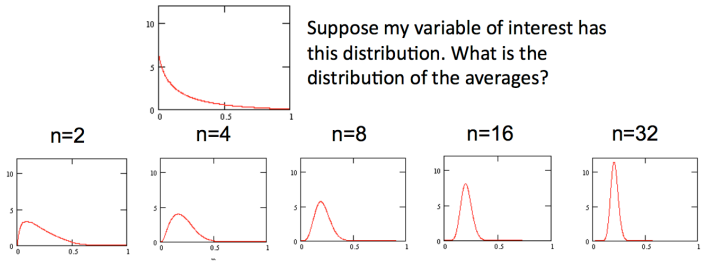
Lecture 3: Hypothesis testing

BMI 713
October 26, 2017
Peter J Park

Review

- For X to be normally distributed, does X have to be normally distributed?
- The distribution of an average tends to be Normal, even when the distribution from which the average is computed is non-Normal (Central Limit Theorem)



A few well-known distributions in statistics

- Normal distribution
- t-distribution
- F-distribution
- χ^2-distribution
- Binomial
- Poisson
- Hypergeometric
- Negative binomial

If $X \sim N(\mu, \sigma^2)$ and $Z = (X - \mu)/\sigma$, then $Z \sim N(0, 1)$.

- In fact, this is the form of most statistical testing:

$$\frac{\text{Statistic - hypothesized value}}{\text{Square root of the variance of the statistic}}$$

follows a known probability distribution

- The probability that the true population mean μ will be contained in $\left(\bar{x} - 1.96\,\frac{\sigma}{\sqrt{n}}\,,\ \bar{x} + 1.96\,\frac{\sigma}{\sqrt{n}}\right)$ is 95%

- What does this mean?
- After we draw the sample, is it correct to say "The probability that μ is contained in the interval is 95%"?
- μ is fixed, not random. Once we have calculated the interval, it simply either contains μ or it doesn't.

Hypothesis Testing

- Specify the null hypothesis (H_0) and the alternative hypothesis (H_1)
- Select a significance level and calculate the statistic
- Calculate the **p-value (the probability of obtaining a statistic as extreme or more extreme under the null hypothesis)**
- Describe the result and the conclusion in an understandable way
- **You "fail to reject H_0" rather than "accept H_0"**

One-sample inference

- **Null hypothesis**: a statement that the population parameter is equal to some particular value of interest.

$$H_0 : \mu = \mu_0$$

- "Proof by contradiction": Null hypothesis is typically what we want to disprove.
- **Alternative hypothesis**:

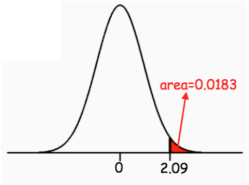
$$H_0 : \mu = \mu_0$$
$$H_1 : \mu < \mu_0 \qquad H_1 : \mu \neq \mu_0 \qquad H_1 : \mu > \mu_0$$

- Strategy: check whether the difference between the sample mean and the "null value" μ_0 is too big to be due to chance alone

Example

- Fasting plasma glucose levels are measured on a sample of 20 mice. The sample average is 107 mg/dL. Suppose that the standard deviation in this population is known to be 15. Is there evidence that this population has average FPG > 100 (i.e., impaired glucose tolerance)?

$$H_0 : \mu = 100$$
$$H_1 : \mu > 100$$
$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{107 - 100}{15/\sqrt{20}} = 2.09$$
$$P(Z \geq 2.09) = 0.0183$$



Interpreting P-values

- Small p-value → data would have been unlikely if H_0 were true, so reject H_0
- If H_0 is rejected, the result is "statistically significant"
- If H_0 is not rejected, it does not mean that H_0 is true.
- **"Not guilty" is not the same thing as "innocent"!**

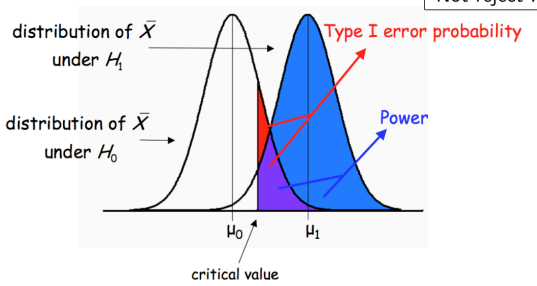
- It is incorrect to talk about the "probability that H_0 is true" (or false). Either it's true or it's not -- we just don't know.
- Inference means deciding whether to believe it's true or not.

Interpreting P-values

- The smaller the p-value, the more convinced we are that it's real.
- **Effect size**: A small p-value does not mean that the difference between μ_0 and the true value μ is large.
- In other words, **statistical significance measures whether a result is "real", not whether it's large**
- Example:
  - With genomic data, it is easy to get a small p-value due to the large number of data points!
  - The correlation coefficient between two variables may only be .01, but it could still be statistically significant (p<.0001)

Type I and Type II errors

	H_0 is true	H_1 is true
Reject H_0	Type I error	Correct
Not reject H_0	Correct	Type II error



- P(Type I error) = P(reject H_0 | H_0 is true) = α ("false alarm")
- P(Type II error) = P(not reject H_0 | H_1 is true) = β ("alarm failure")
- Power = P(reject H_0 | H_1 is true) = 1 – β

Type I and Type II errors

- The probability of committing a type I error is represented by α and is called the significance level of the test
- If α= 0.05 and if repeated tests of hypothesis are conducted based on samples of size *n*, a true null hypothesis would be rejected 5% of the time
- Failing to reject the null hypothesis when it is false is called a type II error
- The probability of making a type II error is denoted by β
- We would like both α and β to be as small as possible

What if variance is unknown?

- What is the problem with the previous examples?
- We do not know the population σ.
- We estimate σ with the sample standard deviation *s*

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

- This introduces another source of uncertainty, so we must modify our hypothesis test to reflect that. This modification changes our "z-test"to a "t-test"

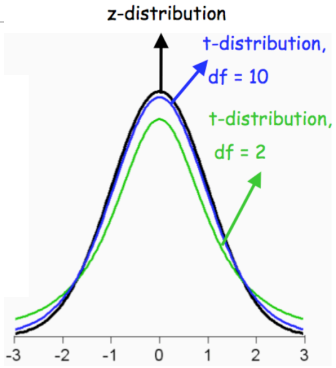
One sample t-test

- z-test

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

- t-test

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$



- t-distribution has fatter tails than z-dist; more diffuse distribution reflects greater uncertainty
- For large *n*, t_{n-1} is nearly identical to the Normal

A little bit of statistical theory

- These are asymptotic results
- That means the result (e.g., p-value) becomes more accurate as the sample size gets large
- How big should my sample size be? How quickly does it become valid?
- It depends on the underlying distribution: if the underlying distribution is normal, then you do not need as many samples
- Most text books will give you guidelines
- What if my sample size is still small?

When the sample size is small

- What is the problem with the t-test in this case?
  - You may have an inaccurate estimate of the variance
- Example from genome-wide gene expression analysis
  - The top genes might be those for which the variance was underestimated due to small sample size
- "Regularized t-test"
  - One solution using a "fudge factor" s_0 (Tusher *et al, PNAS*, 2003)

$$t = \frac{\bar{X} - \mu}{s_0 + s/\sqrt{n}}$$

Some thoughts

- There are many assumptions behind the tests
  - Distributional assumption
  - minimum sample size
- Recognize that every statistic has flaws--the question is whether it is severe enough to invalidate the conclusion
- Consult a statistician for help but recognize that not all statisticians are the same

Permutation test

- A randomized experiment (Box, Hunter, *Statistics for Experimenters*)

Randomized, blocked treatment layout for fertilizer

A	B	B	A	B	
B	A	A	B	A	B

Pounds of tomatoes harvested per plot

11.4	23.7	26.6	21.1	17.9	
28.5	29.9	16.5	24.3	25.3	14.2

Question: Does treatment B provide a better yield?

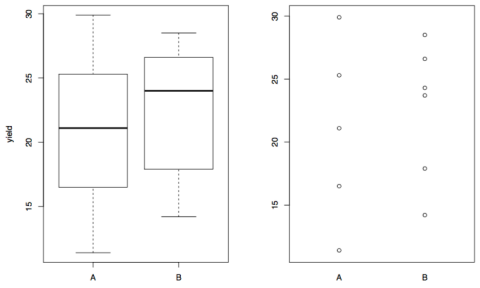
- Question: Does treatment B provide a better yield?
- More formally: Let X denote the random variable governed by the distribution of yields under treatment A, and let Y follow the distribution of yields under treatment B. The question may be restated as: Is E(Y) > E(X)?
- Approach: Randomized application to the plots, blocked east to west, north:south treatment allocations a random permutation of (A,B).
- Let x_i, i = 1,...5 denote yields with treatment A, and y_j, j = 1, . . . , 6 denote yields with treatment B.
- These will be used to test H_0: E(X) = E(Y) vs. H_1: E(Y) > E(X).

A test statistic: the difference in means

- To test the null hypothesis, consider the difference of sample means

$$\bar{\Delta} = \bar{y} - \bar{x}$$

- In this example, the mean difference is 1.693



If you did not know any statistical theory…

- One way of realizing the distribution under the assumption of no effect of treatment on yield is to
  - permute the treatment labels
  - compute the hypothetical mean difference using these newly assigned labels to form the groups
- How many ways are there to assign 5 A and 6 B labels to 11 plots?

R code

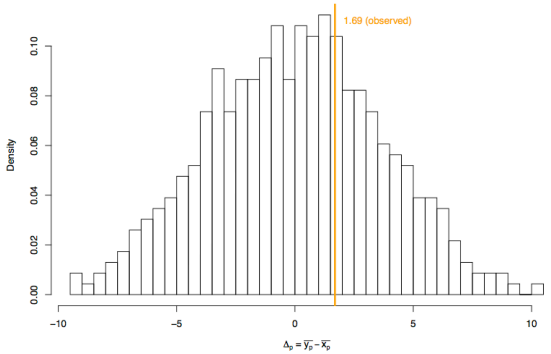
The data:

```
trt = rep(c("A", "B"), c(5,6))
dat = c(29.9,11.4,25.3,16.5,21.1,26.6,23.7,28.5,14.2,17.9,24.3)
names(dat) = trt
dat
```

```
##    A    A    A    A    A    B    B    B    B    B    B
## 29.9 11.4 25.3 16.5 21.1 26.6 23.7 28.5 14.2 17.9 24.3
```

```
Ainds = combn(1:11, 5)
allpd = apply(Ainds, 2, function(x) mean(dat[-x])-mean(dat[x]))
```


The permutation distribution of the differences



- The differences obtained under label permutation are frequently larger than the value observed in the field.

Interpretation

- The data summary $\bar{\Delta} = \bar{y} - \bar{x} > 0$ is consistent with the hypothesis that treatment B is superior to treatment A
- However, this finding might be a manifestation of a “chance fluctuation”
- The distribution of Δ_p obtained when class labels are permuted helps us to understand the scope of variation when the values observed are completely independent of the treatment assignment
- The fact that a substantial fraction (actually 33%) of the Δ_p are larger than the observed $\bar{\Delta}$ suggests that the observation is not particularly unlikely under the assumption that E(Y) = E(X)
- 0.33 is an empirical one-sided p-value for the null hypothesis

Using the t-distribution