EPSE 592: Design & Analysis of Experiments

Ed Kroc

University of British Columbia

ed.kroc@ubc.ca

January 23, 2020

Last Time

- Bernoulli, Binomial, Normal r.v.s
- Sample statistics
- Standard errors

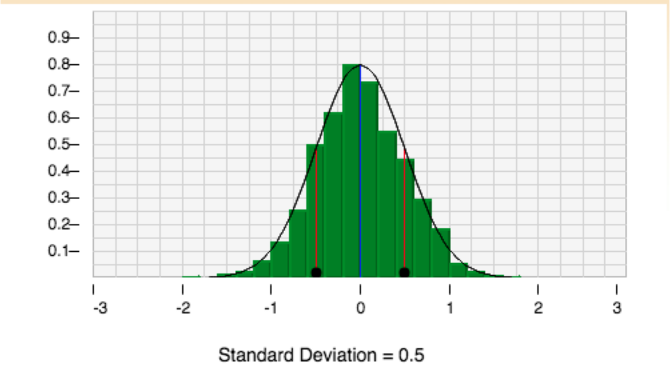
Last Time

- Confidence intervals
- Central Limit Theorem
- Hypothesis testing
- Test statistics and p-values
- Z-test, t-test, F-test

Sample Statistics

- In practice, we study a random variable by observing its values on only a *sample*.
- Studying this sample allows us to infer properties of the actual random variable if the sample is random and representative.
- This is basically what applied statistics is all about!

- We can approximate a r.v.'s PMF or PDF by plotting a *histogram* of our sample data.



Visit: http://www.shodor.org/interactivate/activities/NormalDistribution/

Sample Statistics

- We can get a sense of the "typical" value of our r.v. by calculating a *sample mean*, *sample median*, or *sample mode*.

- Let $\{X_1, \ldots, X_n\}$ denote a random sample of n independent observations from the random variable X. We define the **sample mean** by:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

- sample median = 50th percentile of sample data

- sample mode = most commonly observed value in sample data

- Rememeber: these can all be different!

- We can get a sense of the spread or dispersion (variability) of our r.v. by calculating a *sample variance*.

- Let $\{X_1, \ldots, X_n\}$ denote a random sample of n independent observations from the random variable X. We define the **sample variance** by:

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

Sample Statistics vs. Properties of Random Variables

- Although the definitions of *expectation* and *sample mean*, and of *variance* and *sample variance*, look very similar, they are fundamentally different.
  - Sample mean and variance are *functions of the data/sample*. Different samples will generate different values for sample mean/variance *even if the samples are from the same population*.
  - Expectactions and variances of random variables are idealized quantities. They are inherent properties of the random phenomenon we are studying. We usually cannot calculate them in practice; we can only estimate them via our *sample* approximations.

Standard Errors

- Because sample statistics are random (i.e. not fixed) quantities, they are genuine random variables on their own!
- Thus, they have expectations, variances, std. devs. of their own.
- Terminology: the **standard error** of a sample statistic is simply its standard deviation.
- If T denotes a sample statistic, then we usually write $SE(T)$ to denote its standard error.
- In practice, standard errors are functions of the sample size and the original variability in the population from which we sampled our data.

Confidence Intervals

- A confidence interval is a way of summarizing a sample statistic (e.g. sample mean) and its standard error at once.
- An (approximate) 95% confidence interval for the expectation (population mean), μ_X, of a continuous random variable X from a random sample $\{X_1, \ldots, X_n\}$, for large n, is

$$[\bar{X} - 2 \cdot SE(\bar{X}),\ \bar{X} + 2 \cdot SE(\bar{X})]$$

- Notice, this CI depends on the sample; i.e., it is a statistic.
- **Interpretation:** if we resample 100 times and calculate the 95% confidence interval for each new sample, then approximately 95 of those CIs will contain the true (unknown) population mean.

Central Limit Theorem

Central Limit Theorem (CLT)

Let $\{X_1, \ldots, X_n\}$ denote a random sample of n independent observations from a common distribution with finite mean μ and finite variance σ^2. Recall the sample mean is given by

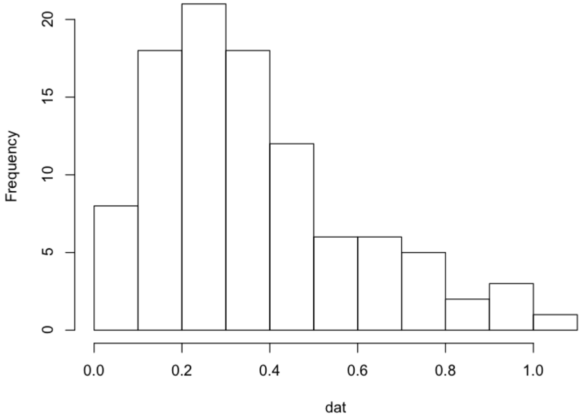
$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Then, for n large, $\bar{X}$ is approximately distributed as $N(\mu, \sigma^2/n)$.

- This is one of the most important theorems of classical statistics. Tells us all about how the sample mean behaves for an independent random sample from *any* common distribution with finite mean and variance.

Central Limit Theorem: example

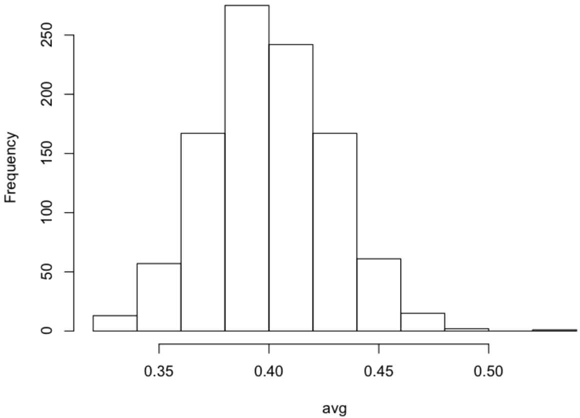
Histogram of **random sample** of size 100 from a very skewed (Gamma) random variable.



Sample mean is 0.366 for this particular set of 100 sample data points.

Central Limit Theorem: example continued

Histogram of the **sample means** of 1000 random samples (each of size 100) from the same very skewed (Gamma) random variable.



Notice the histogram looks quite Normal! (CLT at work)

Central Limit Theorem

- **Moral:** CLT allows us to treat the **sample mean** of *any* random phenomenon as a normal random variable, *as long as our sample size is big enough*.
- This will allow us to assign a measure of uncertainty to our sample mean estimate, e.g. by constructing *confidence intervals*.
- For small sample sizes, either the random phenomenon itself must follow a normal distribution, or we need to use other (nonparametric) statistical methods.

Statistical Hypothesis Testing

- Nearly all quantitative science is based around the idea of stating and testing quantifiable hypotheses about study objects of interest.
- Point Null Hypothesis Testing (PNHT) is the most common option in virtually all applied disciplines.

Statistical Hypothesis Testing

Basic recipe of PNHT:

(1) Identify parameter of interest.

(2) Define null hypothesis, H_0, of *no effect*.

(3) Define a *test statistic* T (a function of the data) such that the larger T is, the less consistent our data are with H_0.

(4) Collect data and then compute test statistic: t_{obs}.

(5) Compute p-value $= \Pr(|T| \geqslant t_{obs} \mid H_0)$; if p-value small enough, then conclude data are **inconsistent** with H_0.

Example:

(1) Difference in mean response between treatment groups X and Y

(2) $H_0 : \mu_X = \mu_Y$

(3) T = standardized difference in sample means

(4) Collect data; compute $t_{obs} = (\bar{X} - \bar{Y})/SE$

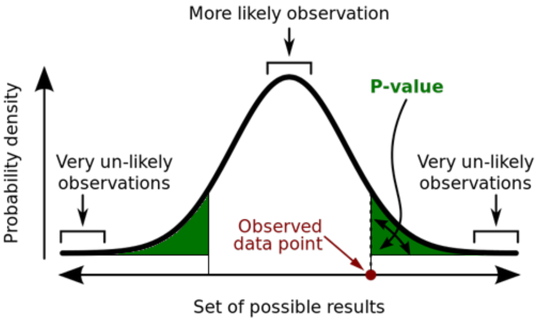
(5) Calculating p-value requires knowing distribution of T given H_0....

A Closer Look at P-values

- Formally, we define

$$\text{p-value} = \Pr(|T| \geqslant t_{obs} \mid H_0), \text{ usually.}$$

- **Interpretation:** the p-value is the probability of observing a test statistic as or more extreme than the one observed for our sample, given that the null hypothesis is true.



A Closer Look at P-values

- Formally, we define

$$\text{p-value} = \Pr(|T| \geqslant t_{obs} \mid H_0), \text{ usually.}$$

- **Interpretation:** the p-value is the probability of observing a test statistic as or more extreme than the one observed for our sample, given that the null hypothesis is true.

- So a big p-value means the observed test statistic is "typical" under H_0. Therefore, the data are consistent with H_0.
- A small p-value means the observed test statistic is *not* "typical" under H_0. Therefore, the data are inconsistent with H_0.

A Closer Look at P-values

- Formally, we define

$$\text{p-value} = \Pr(|T| \geqslant t_{obs} \mid H_0), \text{ usually.}$$

- Recall definition of conditional probability:

$$\Pr(|T| \geqslant t_{obs} \mid H_0 \text{ true}) = \frac{\Pr(|T| \geqslant t_{obs}, \text{ and } H_0 \text{ true})}{\Pr(H_0 \text{ true})}.$$

- With this in mind, how could the p-value be small?

Z-test for Difference of Means

Proposition

If X and Y data come from **normal distributions** *with the* **same known variance** σ^2 *but possibly different means, then the test statistic*

$$T = (\bar{X} - \bar{Y})/SE$$

also follows a normal distribution, with mean $\mu_X - \mu_Y$ and variance σ^2/n, where n denotes the sample size. Therefore, we can calculate

$$p\text{-value} = Pr(|T| \geqslant t_{obs} \mid H_0)$$

since T is $N(0, \sigma^2/n)$ under H_0.

Notice that we do *not* assume anything about μ_X and μ_Y, the quantities we are trying to study. Assuming H_0 (i.e. hypothesis of no difference) allows us to bypass any quantitative assumptions on these parameters.

T-test for Difference of Means

In practice, we are never going to actually know the value of σ^2. Instead, we can estimate it by the *sample variance*. This will allow us to *estimate* the SE.

Proposition

If X and Y are n data points coming from **normal distributions** *with the* **same (unknown) variance** σ^2 *but possibly different means, then the test statistic*

$$T = (\bar{X} - \bar{Y})/\widehat{SE}$$

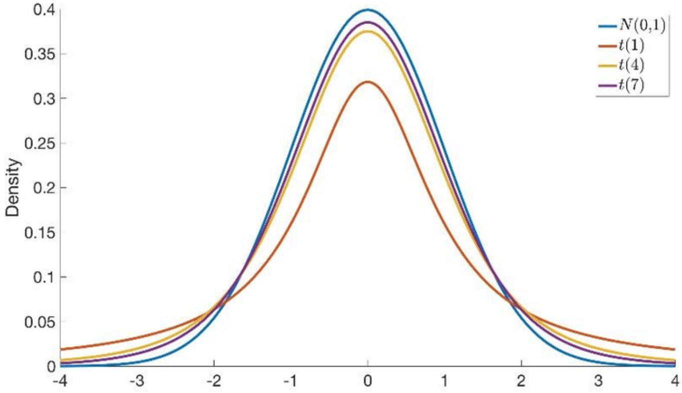
follows a Student-t distribution on $(n-2)$ *degrees of freedom, with mean* $\mu_X - \mu_Y$. *Therefore, we can calculate*

$$p\text{-}value = Pr(|T| \geqslant t_{obs} \mid H_0)$$

since T *is* t_{n-2} *(a known probability disribution) under* H_0.

Student-t Random Variables

- Student-t random variables look like normal distributions, but with *heavy tails*; i.e. extreme events are more likely.



Example: t-test (independent samples)

- Suppose we have annual gross income figures (in $1000's) for a random sample of 10 British Columbians and 10 Albertans:

BC	44	45	46	34	48	42	68	44	52	51
AB	59	50	83	43	65	70	67	77	52	51

- Can use an *independent samples t-test* to test the null hypothesis

$$H_0 \ : \ \mu_{BC} = \mu_{AB}.$$

  Here, we assume that the BC subjects were sampled independently from the AB subjects.

- Must also check assumptions of t-test:
  (1) independence of observations
  (2) normality of data
  (3) homogeneity of variances (homoskedasticity)

Example: t-test (independent samples) in Jamovi

- Enter data as two columns (income, province) in Jamovi

8	44	BC
9	52	BC
10	51	BC
11	59	AB
12	50	AB
13	83	AB
14	43	AB

- Click "Analyses" tab, then "T-Tests", then "Independent Samples T-Test"
- Assign "Income" to dependent variable
- Assign "Province" to grouping variable
- Test statistic, degrees of freedom of (theoretical) Student-t random variable, and p-value will appear in output on right side of screen
- Click on appropriate boxes to produce tests/plots for assumptions, confidence intervals, etc.

Example: t-test (paired samples)

- Suppose we have test scores for 9 first-year calculus students before and after taking a weekend review workshop on pre-calculus topics (algebra, geometry, trigonometry).

before	77	78	82	67	75	91	53	66	70
after	75	80	90	70	70	90	65	74	77

- Can use a *paired samples t-test* to test the null hypothesis

$$H_0 \;:\; \mu_{before} = \mu_{after}.$$

  Here, we are measuring the *same subjects* at two different time points; thus, their responses are **dependent**. A paired t-test accounts for this lack of independence.

- Must also check assumptions of this t-test:
  (1) normality of data

Example: t-test (paired samples) in Jamovi

- Enter data as two columns (before, after) in Jamovi

	before	after
1	77	75
2	78	80
3	82	90
4	67	70
5	75	70

- Click "Analyses" tab, then "T-Tests", then "Paired Samples T-Test"
- Assign "before" and "after" to paired variables
- Test statistic, degrees of freedom of (theoretical) Student-t random variable, and p-value will appear in output on right side of screen
- Click on appropriate boxes to produce tests/plots for assumptions, confidence intervals, etc.

Testing for Equality of Variances

- Equality of variances is an assumption for an *unpaired* t-test.
- But how can we rigorously test if two variances are (statistically) equal?

Sample 1	51	53	49	40	55	56	49	48	42	51
Sample 2	47	45	35	50	70	62	49	37	57	63

- Can calculate sample variances of the two samples: use formula or use "Descriptives" tab in Jamovi.
- $S_1 = 5.15$ and $S_2 = 11.4$
- But are these statistically different? Remember: sample variances are *random variables*. So is this observed difference in sample variances meaningful, given the inherent randomness of the data?

F-test for Inequality of Variances

Proposition

Suppose we draw n_1 sample points from the random variable X and n_2 sample points from the random variable Y. If these X and Y data come from **normal distributions** *with possibly different means and possibly different variances σ_1^2 and σ_2^2, then the test statistic*

$$T = \frac{S_1^2}{S_2^2}$$

follows a Fisher-F distribution on $(n_1 - 1)$ numerator degrees of freedom and $(n_2 - 1)$ denominator degrees of freedom under the null hypothesis

$$H_0 \ : \ \sigma_1^2 = \sigma_2^2.$$

As before, small p-value should reflect when T is an "extreme" value under H_0. This happens if $S_1 >> S_2$ or if $S_1 << S_2$.

Testing for Equality of Variances

- Back to our example:

Sample 1	51	53	49	40	55	56	49	48	42	51
Sample 2	47	45	35	50	70	62	49	37	57	63

- $S_1 = 5.15$ and $S_2 = 11.4$
- In Jamovi, follow the procedure for an independent samples t-test from before.
- Under "Assumption Checks," click the box for "Equality of variances."
- Produces Levene's Test, (essentially) the test statistic S_1^2/S_2^2 compared against its theoretical F distribution under H_0.

F-Tests

- F-tests always take the form of a ratio of variances.
- When the two variances describe normal data, then the ratio of sample variances is a Fisher-F random variable.
- Will rely heavily on this all term: we will usually assume model errors are normally distributed. So can use F-tests to compare if the variance of one model is significantly less than another model (i.e. if one model explains more of the variation in the data than another model).

Summary of statistical tests so far...

- Z-test for testing difference of two group means from (approx.) normal data with known variance.
- T-test for testing difference of two group means from (approx.) normal data with unknown variance. Paired and unpaired versions.
- F-test for testing difference of two group variances from (approx.) normal data.

Note: the CLT implies that we can use *all* these tests for non-normal data as long as we have large enough sample sizes.

Summary of statistical tests so far...

- Z-test for testing difference of **two** group means from (approx.) normal data with known variance.
- T-test for testing difference of **two** group means from (approx.) normal data with unknown variance. Paired and unpaired versions.
- F-test for testing difference of **two** group variances from (approx.) normal data.

Note: the CLT implies that we can use *all* these tests for non-normal data as long as we have large enough sample sizes.

- What about when we want to test for a difference between *more than two* group means?

Example: three experimental groups of interest

Suppose we are interested in studying how amount of higher education correlates with self-reported anxiety levels. We have a survey designed to measure anxiety and give it to 18 people at UBC: 6 who have obtained Bachelor's degrees, 6 who have obtained Master's degrees, and 6 who have obtained PhDs (chosen how?).

Bachelor's	Master's	PhD
6.2	6.2	6.9
5.8	6.9	9.0
6.0	6.2	7.7
5.9	7.7	9.1
6.6	6.8	8.3
6.2	7.9	8.0

Table: Self-reported anxiety levels, 10 point scale. 18 respondents.

Example: three experimental groups of interest

- Could perform 3 independent-samples t-tests to test the 3 null hypotheses:
  - $H_{0,1} : \mu_B = \mu_M$

Independent Samples T-Test

		statistic	df	p
A	Student's t	–2.63	10.0	0.025

  - $H_{0,2} : \mu_M = \mu_P$

Independent Samples T-Test

		statistic	df	p
C	Student's t	2.71	10.0	0.022

  - $H_{0,3} : \mu_B = \mu_P$

Independent Samples T-Test

		statistic	df	p
E	Student's t	–5.73	10.0	<.001

Example: three experimental groups of interest

- Could perform 3 independent-samples t-tests to test the 3 null hypotheses:
  - $H_{0,1} : \mu_B = \mu_M \implies p\text{-value} < 0.05$
  - $H_{0,2} : \mu_M = \mu_P \implies p\text{-value} < 0.05$
  - $H_{0,3} : \mu_B = \mu_P \implies p\text{-value} << 0.05$

- But what about inflated Type I error?

Type I and Type II Errors

- Recall: when p-value small, conclude data inconsistent with H_0.
- Recall: when p-value large, conclude data consistent with H_0.
- Whenever we make a decision about a hypothesis based on a p-value, we have a chance of making an error.

	H_0 true	H_0 false
data inconsistent with H_0	Type I error *false positive*	Correct decision *true positive*
data consistent with H_0	Correct decision *true negative*	Type II error *false negative*

Type I and Type II Errors

- Traditionally, we set a predetermined *significance level*, α, such that

$$\Pr(\text{Type I error}) = \Pr(p - value < \alpha \mid H_0 \text{ true}) = \alpha.$$

- Then α, sample size, variability, and choice of test determine

$$\Pr(\text{Type II error}) = \Pr(p - value > \alpha \mid H_0 \text{ false}) = \beta.$$

- The *confidence level*, or *specificity*, of a test is defined as

$$\Pr(p - value > \alpha \mid H_0 \text{ true}) = 1 - \alpha.$$

- The *power*, or *sensitivity*, of a test is defined as

$$\Pr(p - value < \alpha \mid H_0 \text{ false}) = 1 - \beta.$$

Type I and Type II Errors

- In practice, $\alpha = 0.05$ is a common choice.
- Note: all of $1-\alpha$, β, and $1-\beta$ are determined once α has been fixed, the data have been collected, and the choice of analysis made.
- Good studies will strive to have $1-\beta \geqslant 0.80$. Most studies will have much lower power.

	Given H_0 true	Given H_0 false
Pr(data inconsistent with $H_0 \mid \cdots$)	α	$(1-\beta)$
Pr(data consistent with $H_0 \mid \cdots$)	$(1-\alpha)$	β

Multiple Testing

- Each time we conduct a statistical test of hypothesis, we have a chance of committing a Type I or Type II error.
- The choice of α controls our chance of Type I error for a single test.
- Thus, if our study requires more than one test, each one has a chance of error.
- Thus, if our study requires more than one test, we should be concerned with the *family-wise* error rate: the probability of committing *at least one* Type I error.