Introduction to Statistical Significance

BUS 735: Business Decision Making and Research

1 Goals and Agenda

1.1 Goals

Goals

Re-familiarize ourselves with basic statistics ideas: sampling distributions, hypothesis tests, p-values.

Learning Outcomes

- Background for learning outcomes LO1 and L02 regarding methods of statistical analysis
- LO6: Be able to use standard computer packages such as R to conduct statistical analysis

Agenda

Learning Objective	**Active Learning Activity**
Re-familiarize ourselves with basic statistics ideas: sampling distributions, hypothesis tests, p-values.	Lecture / Discussion
Get comfortable with R environment and programming language	Online Tutorial

2 Statistical Significance

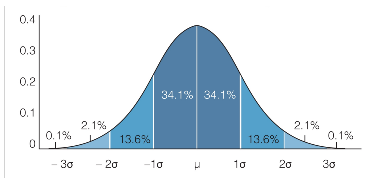
2.1 Sampling Distribution

Probability Distribution

Probability distribution: summary of all possible values a variable can take along with the probabilities in which they occur.

Probability Distribution

Picture



Table

z	0.00	0.01	0.02	0.03	0.04	0.05
0.0	0.5000	0.5040	0.5080	0.5120	0.5159	0.5199
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749

Formula

$$f(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{(x-\mu)^2}{2\sigma^2}\right)}$$

Computer (R example)

```
> pnorm(1.0)
```

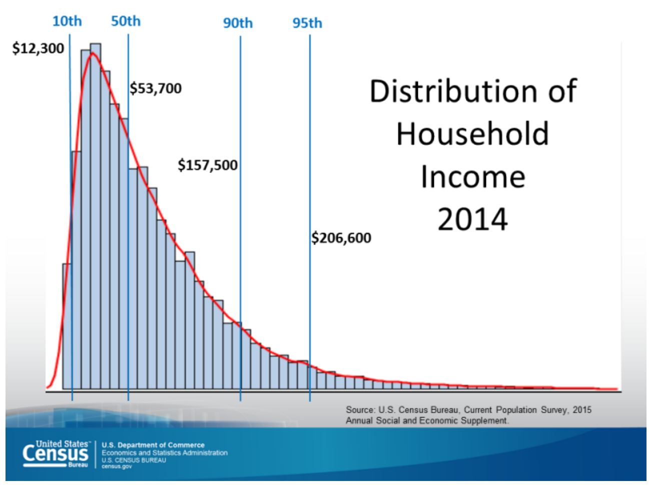
returns $P(z < 1.0) = 0.8413$

Probability Distribution

Probability distributions are typically defined by...

1. Measure of center, such as the `mean` of the distribution
2. Measure of spread, such as the `variance` or `standard deviation`
3. Shape, eg. symmetric, bell-shaped, defined explicitly with an equation

Example: Estimated Income Distribution



Sampling distribution

- Imagine taking a sample of size 100 from a population and computing some kind of statistic.

- The statistic you compute can be anything, such as: mean, median, proportion, difference between two sample means, standard deviation, variance, or anything else you might imagine.
- Suppose you repeated this experiment over and over: take a sample of 100 and compute and record the statistic.
- A **sampling distribution** is the probability distribution *of the statistic*
- Is this the same thing as the probability distribution of the population? NO! They may coincidentally have the same shape though.

Sampling Distribution Simulator

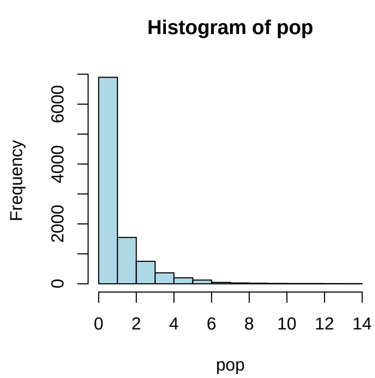
Sampling Distribution Simulator
http://onlinestatbook.com/stat_sim/sampling_dist/

Example in R: Create Hypothetical Population

```
# Generate random pop. w/ 10,000 obs from a Chi-Square dist.
pop <- rchisq(10000, 1)
# Compute the population mean
mean(pop)
## [1] 0.9804668
# Compute the population variance
var(pop)
## [1] 1.873752
# Compute the population std dev
sqrt(var(pop))
## [1] 1.368851
```

Example in R

```
hist(pop,col='lightblue')
```

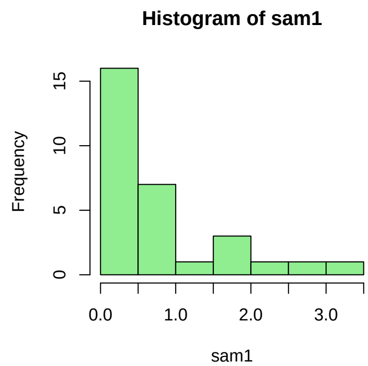


Population mean = 0.98   Population std dev = 1.369   Population is skewed to the right

Generate Samples

```r
# Generate one sample of size 30
sam1 <- sample(pop,30)
mean(sam1)
## [1] 0.7461987
sqrt(var(sam1))
## [1] 0.8968072
```

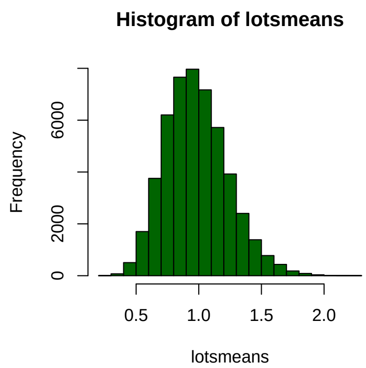
```r
hist(sam1,col='lightgreen')
```



Generate Sampling Distribution

```r
# Generate 50,000 samples,
# Each 30 obs, compute each mean
lotsmeans <-
  replicate(50000,
     mean( sample(pop,30) ) )
# Mean of all the means
mean( lotsmeans )
## [1] 0.9824764
# Std dev of all the means
sqrt(var(lotsmeans))
## [1] 0.2498053
```

```r
# Histogram of all the means
hist(lotsmeans, col='darkgreen')
```



Purpose of a Sampling Distribution

- In reality, you only do an experiment once, so the sampling distribution is a hypothetical distribution.
- Why are we interested in this?

Desirable qualities

What are some qualities you would like to see in a sampling distribution?

- The average of the sample statistics is equal to the true population parameter.
- Want the variance *of the sampling distribution* to be as small as possible. Why?
- Want the *sampling distribution* to be normal, regardless of the distribution of the population.

2.2 Central Limit Theorem

Central Limit Theorem

- Given:
  - Suppose a RV x has a distribution (it need not be normal) with mean μ and standard deviation σ.
  - Suppose a *sample mean* ($\bar{x}$) is computed from a sample of size n.
- Then, if n is sufficiently large, the sampling distribution of $\bar{x}$ will have the following properties:
  - The sampling distribution of $\bar{x}$ will be normal.

- The mean of the sampling distribution will equal the mean of the population (unbiased):

$$\mu_{\bar{x}} = \mu$$

- The standard deviation of the sampling distribution will decrease with larger sample sizes, and is given by:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Central Limit Theorem: Small samples

If n is small (rule of thumb for a single variable: $n < 30$)

- The sample mean is still *unbiased*.
- Formula for the standard deviation of sampling distribution still valid
- Given a small sample size, standard deviation of sampling distribution may be large
- Sampling distribution will be normal *if the distribution of the population is normal*

Example 1

Suppose average birth weight is $\mu = 7lbs$, and the standard deviation is $\sigma = 1.5lbs$.

What is the probability that a sample of size $n = 30$ will have a mean of $7.5lbs$ or greater?

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

$$z = \frac{7.5 - 7}{1.5/\sqrt{30}} = 1.826$$

The probability the sample mean is greater than $7.5lbs$ is:

$$P(\bar{x} > 7.5) = P(z > 1.826) = 0.0339$$

Example 2

Suppose average birth weight is $\mu = 7lbs$, and the standard deviation is $\sigma = 1.5lbs$.

What is the probability that a randomly selected baby will have a weight of 7.5lbs or more? What do you need to assume to answer this question?
Must assume the population is normally distributed. Why?

$$z = \frac{x - \mu}{\sigma} = \frac{7.5 - 7}{1.5} = 0.33$$

The probability that a baby is greater than 7.5lbs is:

$$P(x > 7.5) = P(z > 0.33) = 0.3707$$

Example 3

- Suppose average birth weight of all babies is $\mu = 7lbs$, and the standard deviation is $\sigma = 1.5lbs$.
- Suppose you collect a sample of 30 newborn babies whose mothers smoked during pregnancy.
- Suppose you obtained a sample mean $\bar{x} = 6lbs$. If you assume the mean birth weight of babies whose mothers smoked during pregnancy has the same sampling distribution as the rest of the population, what is the probability of getting a sample mean this low or lower?

Example 3 continued

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

$$z = \frac{6 - 7}{1.5/\sqrt{30}} = -3.65$$

The probability the sample mean is less than or equal to $6lbs$ is:

$$P(\bar{x} < 6) = P(z < -3.65) = 0.000131$$

That is, if smoking during pregnancy actually does still lead to an average birth weight of 7 pounds, there was only a 0.000131 (or 0.0131%) chance of getting a sample mean as low as six or lower. This is an extremely unlikely event if the assumption is true. Therefore it is likely the assumption is not true.

2.3 Hypotheses Tests

Statistical Hypotheses

- A **hypothesis** is a claim or statement about a property of a population.
  - Example: The population mean for systolic blood pressure is 120.
- A **hypothesis test** (or **test of significance**) is a standard procedure for testing a claim about a property of a population.
- Recall the example about birth weights with mothers who smoked during pregnancy.
  - Hypothesis: Smoking during pregnancy leads to an average birth weight of 7 pounds (the same as with mothers who do not smoke).

Null and Alternative Hypotheses

- The **null hypothesis** is a statement that the value of a population parameter (such as the population mean) *is equal to* some claimed value.
  - H_0: $\mu = 7$.
- The **alternative hypothesis** is an alternative to the null hypothesis; a statement that says a parameter differs from the value given in the null hypothesis.
  - H_a: $\mu < 7$.
  - H_a: $\mu > 7$.
  - H_a: $\mu \neq 7$.
- In hypothesis testing, assume the null hypothesis is true until there is strong statistical evidence to suggest the alternative hypothesis.
- Similar to an "innocent until proven guilty" policy.

Hypothesis tests

- (Many) hypothesis tests are all the same:

$$z \text{ or } t = \frac{\text{sample statistic} - \text{null hypothesis value}}{\text{standard deviation of the sampling distribution}}$$

- Example: hypothesis testing about μ:
  - Sample statistic = $\bar{x}$.
  - Standard deviation of the sampling distribution of $\bar{x}$:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

P-values

- Interpretation: *If the null hypothesis is correct,* than the p-value is the probability of obtaining a sample that yielded your statistic, or a statistic that provides even stronger evidence for the alternative hypothesis.
- The p-value is therefore a measure of *statistical significance.*
  - If p-values are small, there is sufficient statistical evidence in favor of the alternative hypothesis.
  - If p-values are large, there is insignificant statistical evidence. Therefore, you **fail to reject the null hypothesis**.
- Best practice is writing research: report the p-value, report your significance level (cut-off value), then reject / fail-to-reject.

3 Using R

Using R

Online tutorial for first-time R user:`http://tryr.codeschool.com/` Other resources:

- *R for Beginners*: PDF manual for learning R `https://cran.r-project.org/doc/contrib/Paradis-rdebuts_en.pdf`
- *An Introduction to R*: PDF Reference Manual for common R tools `https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf`
- Google It! `http://www.google.com`
  - Usually useful results from Stackexchange.com or Stackoverflow.com