

Sampling Distributions & Probability

Paul Gribble

Winter, 2017

McCall Chapter 3

- ▶ measures of central tendency
 - ▶ mean
 - ▶ deviations about the mean
 - ▶ minimum variability of scores about the mean
 - ▶ median
 - ▶ mode

McCall Chapter 3

- ▶ measures of variability
 - ▶ range
 - ▶ variance
 - ▶ standard deviation

Population vs Sample

- ▶ why do we sample the population?
- ▶ in cases when we cannot feasibly measure the entire population
- ▶ the idea is that we can use characteristics of our sample to **estimate** characteristics of the population

McCall Chapter 3

- ▶ populations vs samples
 - ▶ estimators of population parameters
 - ▶ based on a sample
 - ▶ e.g. for estimating parameters of normal distribution
 - ▶ mean, variance

3-4 SUMMARY OF THE NAMES, SYMBOLS, AND FORMULAS FOR COMMON STATISTICS AND PARAMETERS

Quantity	Sample Statistic			Population Parameter		
	Symbol	Read As	Formula	Symbol	Read As	Formula ¹
Mean	\bar{X}	“X bar”	$\frac{\sum X}{N}$	μ	“mew”	$\frac{\sum X}{N}$
Variance	s^2	“s squared”	$\frac{\sum (X - \bar{X})^2}{N - 1}$	σ^2	“sigma squared”	$\frac{\sum (X - \mu)^2}{N}$
Standard Deviation	s	“s”	$\sqrt{\frac{\sum (X - \bar{X})^2}{N - 1}}$	σ	“sigma”	$\sqrt{\frac{\sum (X - \mu)^2}{N}}$

McCall Chapter 7

- ▶ sampling
- ▶ sampling distribution
- ▶ sampling error
- ▶ probability & hypothesis testing
- ▶ estimation

Methods of Sampling

- ▶ simple random sampling
 - ▶ all elements of the population have an equal probability of being selected for the sample
 - ▶ representative samples of all aspects of population (for large samples)

Methods of Sampling

- ▶ proportional stratified random sample
 - ▶ mainly used for small samples
 - ▶ random sampling within groups but not between
 - ▶ e.g. political polls
 - ▶ random sampling within each province
 - ▶ but not between provinces
 - ▶ total # samples for each province pre-determined by overall population

Random Sampling

- ▶ each subject is **selected independently** of other subjects
- ▶ selection of one element of the population does not alter likelihood of selecting any other element of the population

Sampling in Practice

- ▶ elements of the population available to be sampled is often biased
 - ▶ willingness of subjects to participate
 - ▶ certain subjects sign up for certain kinds of experiments
 - ▶ Psych 1000 subject pool — is it representative of the general population?

Sampling Distributions

- ▶ sampling is an imprecise process
- ▶ estimate will never be exactly the same as population parameter
- ▶ a set of *multiple estimates* based on *multiple samples* is called an **empirical sampling distribution**

Sampling Distribution

Definition (sampling distribution)

the **distribution of a statistic** (e.g. the mean) determined on *separate independent samples of size N* drawn from a given population

Empirical Sampling Distribution

7-1 POPULATION DISTRIBUTION OF 20 RAW SCORES, 10 OBSERVED SAMPLE DISTRIBUTIONS ($N = 4$), AND AN EMPIRICAL SAMPLING DISTRIBUTION OF THE MEANS

Population Distribution of Raw Scores		10 Observed Sample Distributions ($N = 4$)	Empirical Sampling Distribution of the \bar{X} 's
6	2	(1, 5, 9, 0)	3.75
9	5	(0, 3, 1, 5)	2.25
0	1	(5, 8, 3, 0)	4.00
3	2	(1, 5, 0, 7)	3.25
1	1	(7, 6, 1, 3)	4.25
5	2	(3, 2, 1, 7)	3.25
7	7	(2, 0, 3, 5)	2.50
7	8	(1, 2, 1, 1)	1.25
1	1	(2, 7, 1, 7)	4.25
3	7	(9, 7, 6, 2)	6.00
$\mu = 3.90, \sigma = 2.88$			Mean of \bar{X} 's = $\bar{\bar{X}} = 3.48$ Standard deviation of \bar{X} 's = $s_{\bar{x}} = 1.31$

Sampling Distributions

- mean, standard deviation and variance in raw score distributions vs sampling distributions:

7-2 TERMS AND SYMBOLS FOR THE MEAN, STANDARD DEVIATION, AND VARIANCE IN DIFFERENT TYPES OF DISTRIBUTIONS

Distributions of Raw Scores	Mean	Standard Deviation	Variance
Sample	\bar{X}	s_x or s	s_x^2 or s^2
Population	μ_x or μ	σ_x or σ	σ_x^2 or σ^2
Sampling Distribution of the Mean	Mean	Standard Error of the Mean	Square of the Standard Error of the Mean
Sample	$\bar{X}_{\bar{x}}$	$s_{\bar{x}}$	$s_{\bar{x}}^2$
Population	$\mu_{\bar{x}}$ or μ	$\sigma_{\bar{x}}$	$\sigma_{\bar{x}}^2$
Equivalences			

$$\mu_x = \mu_{\bar{x}} = \mu \quad s_{\bar{x}} = \frac{s_x}{\sqrt{N}} \text{ and } \sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{N}}$$

Population Estimates

- ▶ by using the mean of a *sample* of raw scores we can estimate both:
 - ▶ mean of *sampling distribution of means*
 - ▶ *mean of population* raw scores
- ▶ we can estimate the standard deviation of the sampling distribution of the means using: $s_{\bar{x}} = \frac{s_x}{\sqrt{N}}$
 - ▶ standard deviation of raw scores in sample divided by the square root of the size of the sample

Standard error of the mean

- ▶ all that's required to estimate it is
 - ▶ standard deviation of raw scores
 - ▶ N (# scores in sample)
- ▶ it represents an estimate of the amount of variability (or sampling error) in means *from all possible samples of size N* of the population of raw scores

Standard error of the mean

- ▶ this is great news, it means that it's **not** necessary to select several samples in order to estimate the population sampling error of the mean
- ▶ we only need 1 sample, and based on its standard deviation, we can compute an estimate of how our estimate of the *mean* would vary *if* we were to repeatedly sample
- ▶ we can then use our estimate $s_{\bar{x}}$ as a measure of the **precision of our estimate of the population mean**

Standard error of the mean

$$s_{\bar{x}} = \frac{s_x}{\sqrt{N}}$$

- ▶ we are dividing by \sqrt{N}
- ▶ thus $s_{\bar{x}}$ (standard error of the mean) is **always** smaller than s_x (standard deviation of raw scores in a sample)
- ▶ said differently: the variability of means from sample to sample will always be smaller than the variability of raw scores within a sample

Standard error of the mean

- ▶ as N increases, $s_{\bar{x}}$ decreases
- ▶ for large samples (large N), the mean will be less variable from sample to sample
- ▶ and so will be a more accurate estimate of the true mean of the population
- ▶ larger samples produce more accurate and more precise estimates

Normal Distribution

- ▶ given random sampling, the sampling distribution of the mean:
 - ▶ is a normal distribution if the population distribution of the raw scores is normal
 - ▶ approaches a normal distribution as the size of the sample increases even if the population distribution of raw scores is *not* normal
- ▶ **Central Limit Theorem**
 - ▶ the sum of a large number of independent observations from the same distribution has, under certain general conditions, an approximate normal distribution
 - ▶ the approximation steadily improves as the number of observations increases

Normal Distribution

- ▶ why do we care about whether populations or samples are normally distributed?
- ▶ all sorts of *parametric* statistical tests are based on the assumption of a particular theoretical sampling distribution
 - ▶ t-test (normal)
 - ▶ F-test (normal)
 - ▶ others. . .
- ▶ assuming an *underlying theoretical distribution* allows us to quickly compute population estimates, and compute probabilities of particular outcomes quickly and easily
- ▶ non-parametric methods can be used in other cases but they are more work

Normal Distribution

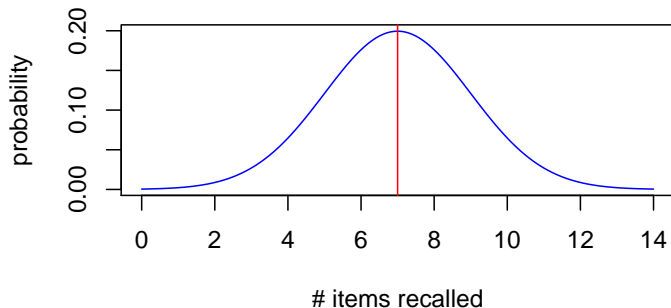
- ▶ given two parameters (mean, variance):
 - ▶ we can look up in a table (or compute in R) the **proportion of population scores that fall above (or below) a given value** (allowing us to compute probabilities of particular outcomes)
 - ▶ we can *assume the shape of the entire distribution* based only on the **mean** and **variance** of our sample

Violations of Normality

- ▶ what if the assumption of normality is violated?
- ▶ we can perform *non-parametric* statistical tests
- ▶ we could determine how serious the violation is (what impact it will have on our statistical tests and the resulting conclusions)
 - ▶ pre-existing rules of thumb about how sensitive a given statistical test is to particular kinds of violations of normality
 - ▶ monte-carlo simulations

A single case

- ▶ suppose it is known:
 - ▶ for a population asked to remember 15 nouns, the mean number of nouns recalled after 1 hour is 7.0, and standard deviation is 2.0 ($\mu = 7.0$; $\sigma = 2.0$)
 - ▶ in R use `dnorm()` to compute probability density

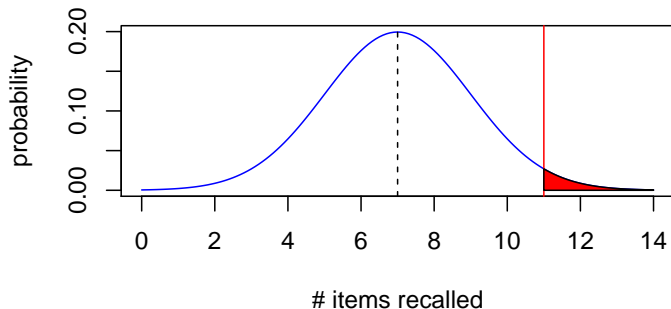


A single case

- ▶ does taking a new drug improve memory?
- ▶ test a single person after taking the drug
- ▶ they score 11 nouns recalled
- ▶ what can we conclude?

A single case

- ▶ 11 nouns recalled after taking drug
- ▶ what are the chances that someone **randomly sampled from the population** (without taking the drug) would have scored 11 or higher?
- ▶ this probability equals the area under the curve:



A single case

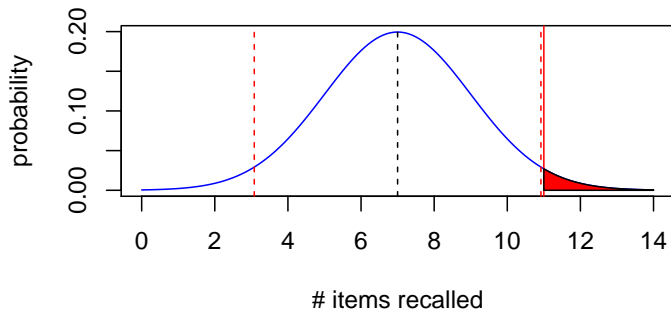
- ▶ to determine probability:
 - ▶ convert score to a *z-score* and lookup in a table
 - ▶ $z = (11.0 - 7.0)/2.0 = 2.0$
 - ▶ or compute directly in R the probability

```
pnorm(11, mean=7, sd=2, lower.tail=FALSE)
```

```
0.0227501319481792
```

A single case

- ▶ $p = 0.0228$ but what is our α level?
- ▶ let's say 5%
- ▶ if we didn't *in advance* have a hypothesis about whether drug should raise or lower memory score, then we need to split our 5% into an upper and lower half:



A single case

- ▶ $p = 0.0228$ and $\alpha = 0.0250$ (two-tailed)
- ▶ thus $p < \alpha$ and so we can reject H_0
- ▶ remember H_0 is that:
 - ▶ the drug has no effect
 - ▶ any difference in our observed sample (in this case 1 score) from the population mean, is **not** due to the drug, but is due to *random sampling error*
 - ▶ i.e. we just happened to randomly sample a person from the population who has good memory
 - ▶ after all the population scores are distributed (normally), some are high, some are low, most are in the middle around 7.0

A single group

- ▶ in this example, mean μ and standard deviation σ of population were known
- ▶ typically we do not know these quantities, and we have to *estimate them from our sample data*

Tests based on estimates: mean

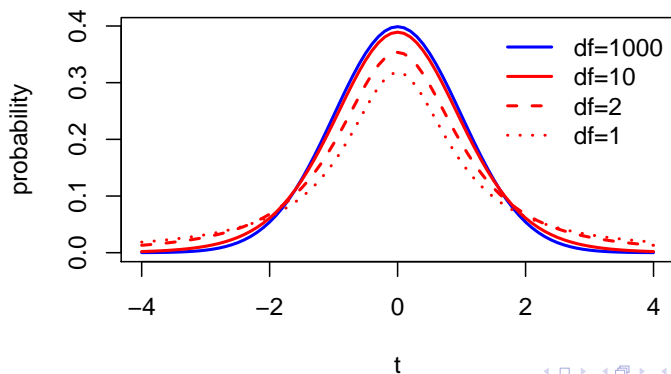
- ▶ it turns out that the best estimate of the population mean μ is the sample mean \bar{X}
- ▶ easy

Tests based on estimates: standard deviation

- ▶ we can use the **standard error of the sampling distribution of the mean** to estimate σ , the standard deviation of the population
- ▶ accuracy of this estimate depends on the sample size N
- ▶ for large samples ($N > 50$, $N > 100$) it's fairly accurate
- ▶ for smaller samples it is not
- ▶ another theoretical sampling distribution exists that is more appropriate for smaller (realistic) sample sizes: **the t distribution**

The t distribution

- ▶ similar to normal (z) distribution
- ▶ however: there is a different shape for each sample size N
- ▶ t distribution characterized by degrees of freedom
 $df = N - 1$



The t Distribution

- ▶ let's sample $N = 20$ subjects at random and give them our memory drug
- ▶ assume population parameter $\mu = 7.0$ and σ is unknown
- ▶ assume scores in population are normally distributed
- ▶ let's test the hypothesis H_0 that the drug has no effect
- ▶ i.e. that the sample is drawn from the population
- ▶ i.e. that any difference between sample and population is due not to the drug, but due to random sampling error

The t Distribution

- ▶ let's say our sample mean is $\bar{X} = 8.4$ and $s = 2.3$
- ▶ compute the t statistic:

$$t_{obs} = (8.4 - 7.0) / (2.3 / \sqrt{20}) = 2.72$$

- ▶ compute the probability of obtaining a t_{obs} this large or larger **under the null hypothesis**

```
pt(2.72, 19, lower.tail=FALSE)
```

```
0.00679475335292515
```

- ▶ since $p < \alpha$ (if we set $\alpha = 0.05$) we can **reject the null hypothesis**
- ▶ we would conclude that we have good evidence that the drug had an effect

Confidence Interval for the mean

- ▶ our sample mean is not equal to the population mean
- ▶ it is an *estimate*
- ▶ using standard error of the mean, and our observed t statistic, we can compute a **confidence interval** for the true population mean

$$\bar{X} \pm t_{\alpha}(s_{\bar{X}})$$

- ▶ in our case:
 - ▶ let's compute the 95% CI (2-tailed)
 - ▶ so $t_{\alpha=.025, df=19} = 2.093$ (use the `qt()` function in **R**)
 - ▶ $8.4 \pm (2.093)(2.3/\sqrt{20}) = (7.33, 9.47)$

Confidence Interval for the mean

- ▶ what does 95% refer to exactly?
- ▶ common misconception: it does **not** mean that there is a 95% chance that the given confidence interval contains the true population mean
- ▶ too bad, this would be a useful thing to know
- ▶ what it **does** mean, is something quite strange:
 - ▶ if we repeatedly sample from the population, each time with sample size N , and for each sample compute its own 95% confidence interval, then 95% of those confidence intervals will contain the true population mean
- ▶ less useful but it's the truth

t-tests for the difference between means

- ▶ assume we have **two** random samples
- ▶ we want to test whether these two samples have been drawn from:
 - ▶ H_0 : the same population (with the same mean)
 - ▶ H_1 : two populations with different means
- ▶ compute the t statistic according to:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_{\bar{X}_1 - \bar{X}_2}}$$

t-tests for the difference between means

- ▶ under H_0 , $\mu_1 = \mu_2$

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_{\bar{X}_1 - \bar{X}_2}} = \frac{(\bar{X}_1 - \bar{X}_2) - 0}{s_{\bar{X}_1 - \bar{X}_2}} = \frac{(\bar{X}_1 - \bar{X}_2)}{s_{\bar{X}_1 - \bar{X}_2}}$$

- ▶ the numerator terms can be easily computed based on our samples
- ▶ the denominator term can be estimated from our sample data
- ▶ it turns out this denominator, *the standard error of the difference between means*, is estimated differently depending on whether scores in the two samples are **correlated** or **independent**

Independent groups t-test

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left[\frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2} \right] \left[\frac{1}{N_1} \frac{1}{N_2} \right]}}$$
$$df = N_1 + N_2 - 2$$

Correlated groups t-test

- compute D_i as the difference between pairs of scores in each group, then

$$t = \frac{\sum D_i}{\sqrt{\frac{N \sum D_i^2 - (\sum D_i)^2}{N-1}}}$$
$$df = N - 1$$

t-tests in R

- ▶ in R use the `t.test()` function with the `paired=TRUE` or `paired=FALSE` parameter to indicate correlated or independent groups

Interpretation of Statistical Significance

- ▶ statistical "significance" and scientific significant are **not** the same thing
- ▶ if N is large you might find a *statistically significant* difference between groups, that is in fact **tiny** and is **meaningless scientifically**
- ▶ if N is small, you might falsely conclude based on statistical tests that show *no significant difference between groups* that the observed difference between groups is *not significant* even though it may be in fact very large, and very important scientifically

Interpretation of Statistical Significance

- ▶ we should all agree to stop saying *statistically significant* and instead say **statistically reliable**
- ▶ difference between groups is **reliable** not (necessarily) *significant*

Interpretation of Statistical Significance

- ▶ imagine an IQ experiment where $N = 10,000,000$ and $p < 0.000001$
 - ▶ less than 1 in 1 million chance of observing such a difference between groups, due to sampling error alone
- ▶ but what if $\bar{X}_1 - \bar{X}_2$ is just 1.0?
 - ▶ population IQ by definition is $\mu = 100$ and $\sigma = 15$
- ▶ this is in fact a tiny difference in IQ (just 1 point)
- ▶ it appears to be so highly *statistically significant* because N is so large.
- ▶ What we should in fact say is that the difference between groups is **extremely reliable**
- ▶ We should not say that it is "extremely significant"