# STAT22000 Autumn 2013 Lecture 19

Yibi Huang

November 11, 2013

6.2 Test of Significance



## Outline

- null hypothesis, alternative hypothesis
- $z$-statistic
- $P$-value
- significance levels



## Example: Annual Fees

A bank wonders whether waiving the annual credit card fee for customers who charges at least \$3,000 in a year would increase the amount charged on its credit card. To test this, the bank makes this offer to 100 randomly selected customers from its existing 1,000,000 credit card holders (Let's assume that it is a SRS). It then compares how much these customers charge this year with the amount that they charged last year. The mean increase is \$565, and the SD is \$267.

Does the no-fee offer increase the amount charged on the credit card?



## Making the Question More Precise

- **Population**: the existing 1,000,000 credit card holders
- **Parameter**: the population mean $\mu$
  – the difference between the amount charged to the card this year and last year, averaged over the 1,000,000 card holders
- **Sample**: the 100 selected customers
- **Statistic**: the sample mean
  – the difference between the amount charged on the card this year and last year, averaged over the 100 selected customers, $=$ \$565.
- Is the amount charged to the card, averaged over the 1,000,000 card holders, increased?
- In other words, is the the population mean $\mu > \$0$?
- **Null hypothesis (H$_0$)**: $\mu = \$0$ (or $< \$0$).
- **Alternative hypotheses (H$_A$)**: $\mu > \$0$.



## Model for a Simple Random Sample (1)

- For the $N = 1{,}000{,}000$ card holders, let $x_i$ be the difference between the amount charged on the credit card of the $i$th card holder this year and last year, $i = 1, 2, \ldots, N$.
- The population mean $\mu$ is the average of all $x_i$'s

$$\mu = \frac{1}{N}(x_1 + x_2 + \ldots + x_N)$$

- The simple random sample $X_1, X_2, \ldots, X_{100}$ is like 100 draws made at random **without** replacement from these $N = 1{,}000{,}000$ numbers $(x_1, x_2, \ldots, x_N)$
- What is the distribution of one observation $X_j$ in a SRS? See the next slide



## Model for a Simple Random Sample (2)

What is the distribution of $X_j$? Let's first suppose that all $x_i$'s are all different.

- The distribution of $X$ is $P(X_j = x_i) = \dfrac{1}{N}$ for all $i = 1, \ldots, N$.
- Then the expected value of $X_j$,

$$\mathbb{E}(X_j) = x_1 \cdot \frac{1}{N} + x_2 \cdot \frac{1}{N} + \cdots + x_N \cdot \frac{1}{N} = \frac{1}{N}(x_1 + x_2 + \ldots + x_N)$$

  is exactly the population $\mu$
- The variance of $X_j$ is

$$\mathrm{Var}(X_j) = (x_1 - \mu)^2 \cdot \frac{1}{N} + (x_2 - \mu)^2 \cdot \frac{1}{N} + \cdots + (x_N - \mu)^2 \cdot \frac{1}{N} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

  which is called the **population variance**, denoted as $\sigma^2$. Note the sample variance is divided by $n - 1$ (here $n$ is the sample size), but the population variance is divided by $N$.

## Model for a Simple Random Sample (3)

If $x_i$'s are NOT all different, then the distribution of $X_j$ becomes

$$P(X_j = x) = \frac{\#\text{ number of } x_i\text{'s that equal } x}{N} \quad \text{for all } x$$

Nonetheless, it is still true that

$$\mathbb{E}(X_j) = \frac{1}{N}(x_1 + x_2 + \ldots + x_N) = \mu$$
$$\text{Var}(X_j) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 = \sigma^2$$

---

When the sample size $n$ is small relative to the population size $N$, then the observations $X_1, X_2, \ldots, X_N$ in the SRS are nearly independent, i.e., they are **i.i.d., with population mean $\mu$, and population variance $\sigma^2$**.

By CLT, when the sample size $n$ is large, the sample mean $\overline{X}$ has an approximate normal distribution $\overline{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.



## Back to the Annual-Fee Example

- $H_0$: $\mu = \$0$ (or $\mu \leq \$0$.)
- $H_A$: $\mu > \$0$.
- In plain words,
  - $H_0$ means the no-fee offer made no change in the average amount , and
  - $H_A$ means flex-time made some change (up or down) in the average absenteeism.



## Test Statistic

- Suppose we want to test the hypothesis that $\mu$ has a specific value:
$$H_0 : \mu = \mu_0$$
- Since $\overline{X}$ estimates $\mu$, the test is based on $\overline{X}$, which has a (approximately) Normal distribution. Thus,
$$z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$$
has a standard normal distribution, **under the null hypothesis $H_0$**.
- **We use $z$ as the test statistic.**
- For the annual-fee example, $\mu_0 = \$0$, $\overline{X} = \$565$, $\sigma$ is assumed to be known to be \$267. So the $z$-statistic is
$$z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{\$565 - \$0}{\$267/\sqrt{100}} = 21.26$$



## $P$- value

If $H_0$ is right, since the sample size is large, the probability histogram of the sample mean is nearly normal. The probability of getting a value of the $z$ statistic at least as extreme as 21.16 is $1.1 \times 10^{-99}$


- If the average amount charged hasn't changed from last year, only 11 in $10^{100}$ studies similar to this one would have shown a greater apparent change.

This probability is called the $P$-**value**.



## $z$-Statistic

In general, if one wants to test whether the population mean $\mu$ of $n$ i.i.d. observation $X_1, X_2, \ldots, X_n$ (like SRS) is a certain given value $\mu_0$ or not, one can find the $z$-**statistic**

$$z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$

in which $\overline{X}$ is the sample mean $\frac{1}{n}(X_1 + X_2 + \cdots + X_n)$, and $\sigma$ is the **known** SD of $X_i$'s.

The bigger the $z$-**statistic** is, the less consistent the data is with $H_0$, and the more consistent it is with $H_A$.



## $P$-value

The $P$-**value** is the **probability** of getting a result as deviant (or even more so) as the one actually observed, **assuming $H_0$ is true**.

The smaller the $P$-value, the stronger the evidence against $H_0$, the harder to believe $H_0$ is right.

- A $P$-value of 0.0001 means that, if $H_0$ is true, only 1 in 10000 similar experiments would give a result at least as extreme as the one in hand. $\Rightarrow$ That's strong evidence against $H_0$.
- A $P$-value of 1/4 means that, if $H_0$ is true, 1 out of 4 similar experiments would give a result at least as extreme as this one. $\Rightarrow$ No reason to disbelieve $H_0$.

The $P$-value is a measure of how surprising the observed data is, if $H_0$ is true. To put it another way, the $P$-value is a measure of how plausible $H_0$ is, in the light of the observed data.

## Conclusions of the Annual Fee Experiment

- Average amount charged did increase this year
- That's all the test of significance tells you. But the bank would want to know more — how big was the change in average amount charged (for all 1,000,000 card holders)?
  - 95%-confidence interval for the change is $\$565 \pm (2 \times \$21.16) \Rightarrow \$522.68$ to $\$607.37$.
- How important is the change in average amount charged?
  - That's question for the bank, not statistics.
- Did the no-fee offer cause the change?
- Can you suggest a better experimental design?



## Steps in Making a Test of Significance:

- Formulate the *null hypothesis* $H_0$ (and perhaps the *alternative hypothesis* $H_A$ as statements about a parameter for the data (and perhaps the *population* and *parameter* if applicable).
- Define a *test statistic* to measure the difference between the data and what's expected under $H_0$.
- Compute the *P-value* — the probability of getting a value for the test statistic as extreme, or more so, than the one observed.

Exercise: Suppose the bank said the average yearly increment in the amount charged should be at least \$500 to compensate the loss for waiving the annual fee. Can you test whether the the increment is at least \$500?

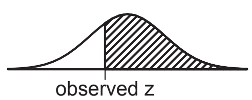


For the one-tailed test with $H_0$ and $H_A$ as follows,

$$H_0 : \mu = \$0,$$
$$H_A : \mu > \$0,$$

if the $z$-statistic is negative, the $P$-value will then be at least 50%, which make sense since in that case, $H_0$ would be more likely to be true than $H_A$.

$P$-value =





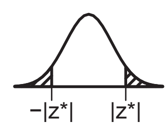

## Two-Tailed Tests v.s. One-Tailed Tests

If the bank think that the no-fee offer could change the amount charged, then $H_A$ can be phrased as

$$H_A\text{: } \mu \neq \$0 \text{ .}$$

In this case, large positive and large negative values of the $z$-statistic are both evidence against $H_0$, and hence the $P$-value is the probability of getting a $z$-statistic with absolute value $\geq$ the observed one ($z^*$)

$P$-value =




Tests with such alternatives are called **two-tailed** (or **two-sided**) tests, and the corresponding $P$-values are called **two-tailed** (or **two-sided**) $P$-values, as oppose to the one-tailed test and one-tailed $P$-value on page 10.



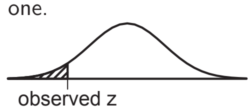
## Another One-Tailed Test

If management believes that no fee offer might decrease the amount charged (which sounds counterintuitive for this example), then $H_A$ will be phrased as

$$H_A\text{: } \mu < \$0,$$

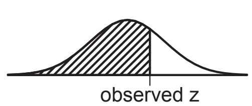
and the $P$-value is the probability of getting a value of the $z$-statistic $\leq$ the observed one.

$P$-value =



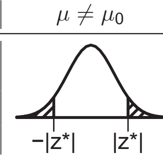
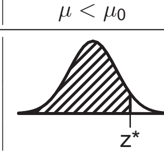
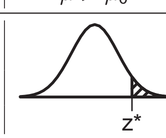
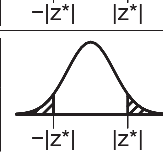
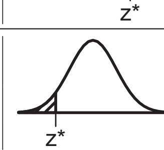
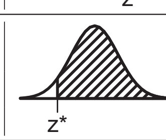

If the $z$-statistic is positive, the $P$-value will then be at least 50%, which make sense since in that case, $H_0$ would be more likely to be true than $H_A$.

$P$-value =






## Summary of the One-Tailed and Two-Tailed Tests

| | Two-tailed test | One-tailed test | |
|---|---|---|---|
| $H_0$ | $\mu = \mu_0$ | $\mu = \mu_0$ or $\mu \geq \mu_0$ | $\mu = \mu_0$ or $\mu \leq \mu_0$ |
| $H_A$ | $\mu \neq \mu_0$ | $\mu < \mu_0$ | $\mu > \mu_0$ |
| $P$-value when $z^* > 0$ |   |   |   |
| $P$-value when $z^* < 0$ |   |   |   |

Here $\mu_0$ is a given value, and $z*$ is the observed $z$-statistic.

## Descriptive v.s. Decision-Theoretic Testing

### Descriptive testing

Statisticians just report the $P$-value(s), and let clients make their own conclusions about the validity of $H_0$.

| $P$-value | Strength of the evidence against $H_0$ |
|---|---|
| near 0 | strong |
| near 1 | weak |

### Decision-theoretic testing

Sometimes, a decision must be reached on the basis of a set of data to reject $H_0$ or not. A common practice is to setup a threshold,

- If $P$-value < threshold, then *reject* $H_0$.
- Otherwise, do not reject $H_0$ ($\neq$ accept $H_0$).

The threshold is called the **level of significance**.



## Significance Levels

- If the $P$-value of a result is less than $\alpha$, the result is said to be **statistically significant at level** $\alpha$, and $H_0$ is rejected at level $\alpha$.
- E.g., a $P$-value of 3.1% is significant at level 5%, but not significant at level 1%.
- Commonly used significance levels: $\alpha = 1\%$ or $\alpha = 5\%$
- If $H_0$ is true, then $H_0$ is rejected at level 0.05 in about only 5 out of every 100 cases.
- If $H_0$ is true, then $H_0$ is rejected at level 0.01 in about only 1 out of every 100 cases.



## True or False

- A result significant at 1% level is real.
  __False__. Even if $H_0$ is true, 1% of the time the experiment will give a result which is "highly significant."

- If a difference is "significant at 1% level," there is less than a 1% probability for $H_0$ to be true.
  __False__. A $P$-value does not give the probability of $H_0$ being true. In fact, the $P$-value is computed assuming $H_0$ is true.



## Analogies between Hypothesis Testing and Criminal Trials

| Criminal Trial | Hypothesis Testing |
|---|---|
| The defendant is innocent | The null hypothesis |
| Verdict of guilty<br>Verdict of not guilty | Reject the null<br>Not reject the null |
| Convicting the innocent | Rejecting $H_0$ when $H_0$ is true. This is at least an embarrassment. You've proclaimed some result is real, but nobody can replicate your findings. A serious set-back to your career. |
| Letting the guilty go free | Accepting $H_0$ when $H_0$ is false. You failed to discover something that's really there. A disappointment to you, but not a setback to science — since it's really there, somebody will find it. |
| Shadow of a doubt<br>Beyond a shadow of a doubt | Significance level<br>$P$-value < significance level |